Final Report

Title: Chance Discovery with Data Crystallization

A Basic Research for Discovering Unobservable Events

Contract Number: FA5209-05-P-0259

AFOSR/AOARD Reference Number: AOARD-05-15

AFOSR/AOARD Program Manager: Tae-Woo Park, Ph.D.

Period of Performance: 01 April 2005 - 31 March 2006

Submission Date: 10 May 2006

PI: Yukio Ohsawa / University of Tsukuba

3-29-1 Otsuka, Bunkyo-ku, Tokyo

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
08 AUG 2006	**Final Report (Technical)**	**01-04-2005 to 31-03-2006**

4. TITLE AND SUBTITLE
Chance Discovery with Data Crystallization - Discovering Unobservable Events

5a. CONTRACT NUMBER
FA520905P0259

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Yukio Ohsawa

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Tsukuba,3-29-1 Otsuka, Bunkyo,Tokyo 112-0012,Japan,JP,1120012

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
The US Resarch Labolatory, AOARD/AFOSR, Unit 45002, APO, AP, 96337-5002

10. SPONSOR/MONITOR'S ACRONYM(S)
AOARD/AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AOARD-054016

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

It is only the observable part of the real world that can be presented in data. For such a scattered, i.e., an incomplete and ill-structured data, data crystallizing aims at presenting the hidden structure by inserting dummy items corresponding to unobservable, i.e., hidden events, to the given data on past events. The existence of hidden events and their position in the environment will be visualized as a result of data crystallizing. This basic method is expected to be applicable for various real world domains to which chance-discovery methods have been applied. This project aims at developing the process of data crystallizing, with a new tool extending KeyGraph, based on the process of chance discovery. In the research, experiments will be made using artificial data obtained from simulating the target of intelligence analysis, i.e., organized crimes.

15. SUBJECT TERMS
Data Mining, Chance Discovery

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**47**	

(2) Objectives: Briefly summarize the objectives of the research effort or the statement of work.

It is only the observable part of the real world that can be presented in data. For such a scattered, i.e., an incomplete and ill-structured data, *data crystallizing* aims at presenting the hidden structure by inserting dummy items corresponding to unobservable, i.e., hidden events, to the given data on past events. The existence of hidden events and their position in the environment will be visualized as a result of data crystallizing. This basic method is expected to be applicable for various real world domains to which chance-discovery methods have been applied. This project aims at developing the process of data crystallizing, with a new tool extending KeyGraph, based on the process of chance discovery. In the research, experiments will be made using artificial data obtained from simulating the target of intelligence analysis, i.e., organized crimes. Then, the method will be applied to real workplaces with real data, real analysts, in real world domains.

(3) Status of effort: A brief statement of progress towards achieving the research objectives. (Limit this section to about 200 words or less.)

*The basic procedure of data crystallizin*s has got realized with a tool which insert dummy items, corresponding to unobservable events, to the given data on past events. The existence of these unobservable events and their relations with other events are visualized by applying KeyGraph iteratively to the data donated with dummy items, gradually increasing the number of edges in the graph, like the crystallization of snow with gradual decrease in the air temperature. For tuning the granularity level of structure to be visualized, this tool is integrated with human's process of chance discovery. Then, a new technique has been developed to understand dark events and to extend the chance discovery process. The technique is human-interactive annealing for revealing latent structures along with the algorithm for discovering dark events. Test data generated from a scale-free network shows that the precision of the algorithm is up to 90%. An experiment on discovering an invisible leader hidden under an on-line decision-making circumstance showed a significantly high performance of the method, and a trial for the analysis on unknown emerging technology has been demonstrated.

(4) Abstract: Briefly describe research accomplishments, their significance to the field, and their relationship to the original goals.

Accomplishments

a. *Stage 1) Development of basic tool* : For a scattered, i.e., an incomplete and ill-structured dataset, we realized a tool for *data crystallizing* which inserts dummy items, corresponding to unobservable events. The existence of these unobservable events and their relations with other events are visualized by applying KeyGraph iteratively to the data donated with dummy items, gradually increasing the number of edges in the graph, like the crystallization of snow with gradual decrease in the air temperature. For tuning the granularity level of structure to be visualized, this tool is integrated with human's process of chance discovery. This basic method came to be proven applicable for the discovery of hidden leaders of meetings, i.e., managers who do not appear in the meeting room but are sending commands to the members who appear in the meetings.

b. *Stage 2) Refinement of the method by weighing human's role in the process of discovery* : We addressed hidden structure visualization adaptive to human's prior understanding. Visualization can be adjusted based on the degree of the user's prior understanding of the problem domain. The degree is represented by a temperature parameter used in the human-interactive annealing along with stable deterministic crystallization algorithm. When the understanding of the problem is believed to be richer, the temperature shall be set higher. More complex higher-order hidden structures shall be revealed. This will lead to the discovery of unique and unexpected scenario. On the other hand, when the understanding is poorer, the temperature shall be set lower. The user should try to understand the basic lower-order structures from the event graph. Such adaptive nature is convenient to discover unexpected scenarios in the individual user's own perspective. The adaptive nature of the annealing process was demonstrated for examples of social network visualizations from: (1) Test data generated from a scale-free network, resulting in the discovery precision of up to 90%. (2) Real on-line communication where people met for group decision, resulting in precisely discovering real leaders who had been deleted from the data of communication (3) data of persons related to famous politicians.

Significance to the field

The basis of this proposal has been *chance discovery*, which means to discover a *chance*, defined as an event significant for making a decision. Using existing data in business and natural/social sciences, we have been achieving successful chance discoveries in various domains, including (not restricted to):

- Marketing, where consumer-behaviors from hidden motivations are dealt with,
- Prediction of earthquakes caused by hidden active faults
- Hepatitis treatment, where some observation might be missing in the blood test.

In studies on chance discovery, we have been working well in finding rare but significant events. Data crystallizing means to extend chance discovery to the discovery of significant events which have never occurred in the given data, i.e., from low-frequency to zero-frequency. This means to deal with more uncertain environment where human may miss important event, than we have been dealing with in data mining or chance discovery.

A relevant research area to Chance Discovery is Evidence Extraction and Link Discovery (EELD), where important links of people with other people and with their own actions are to be discovered from heterogeneous sources of data. The difference between Chance Discovery and EELD, at the time we began this project, was in the position of human factors in the research approaches. In Chance Discovery, the visualization techniques such as KeyGraph have been used for clarifying the effect of chances, by enforcing the user's thoughts on scenarios in the real environment. On the other hand, the EELD program mainly contributed to identifying the most significant links among items more automatically and precisely than human. After the one year of this successful project, we showed an improvement of the visualization tool reinforces the process of chance discovery, and this may be regarded as a new feature of the state of chance discovery.

I expect these two will meet, because the studies in EELD is now oriented to coupling symbolic expressions of human knowledge with a machine learning system. That is, human's interaction with machine intelligence is coming to the centers of these two domains. Some studies in EELD, such as data visualization for decision making, serve bridges between human and machine. In this sense, our methods for data crystallization is expected to contribute to EELD as well as to chance discovery.

Relation to the goal

The sphere of real world applications linked from this basic research is expected to include intelligence analysis aiming to arrest unknown leaders, development of new (unknown) products, aiding corporate behaviors by detecting unknown interest of employees, etc. We successfully accomplished to show the potential ability of our methods to solve these new problems, by applying to toy (simulated) and real problems corresponding to small-size version of these up-to-date problems.

(5) Personnel Supported: List the professional personnel supported by the contract and/or the personnel who participated significantly in the research effort.

Yuki Nyu: Organized the message board where various decision making by a group of 10 to 30 people were made. Significant experimental results have been obtained from her organizational efforts.
Yoshiharu Maeno, Mr: Developed and implemented the new method human-interactive annealing.
Kataichi Ito, Mr: Implemented the basic tool for the experiments of data crystallization

(6) Publications: List peer-reviewed publications submitted and/or accepted during the contract period.

Yoshiharu Maeno and Yukio Ohsawa, Human-Computer Interactive Annealing for Discovering Invisible Dark Events, submitted to IEEE Transaction on Humatronics (Under review 2006)
Yoshiharu Maeno and Yukio Ohsawa, Understanding of dark events for harnessing risk, Chance Discovery for Real World Decision Making, Chapter 22, Springer Verlag (2006)
Kenichi Horie, Yukio Ohsawa, Product Designed on Scenario Maps Using Pictorial KeyGraph, WSEAS Transaction on Information Science and Application, Vol.3 No.7, pp.1324-1331 (2006)
Tsuneki Sakakibara, Yukio Ohsawa, Gradual-Increase Extraction of Target Baskets as Preprocess for Visualizing Simplified Scenario Maps by KeyGraph, Journal of Soft Computing (2006) To Appear
Naohiro Matsumura, Yukio Ohsawaa, Mitsuru Ishizuka, Combination Retrieval for Creating Knowledge from Sparse Document-Collection, Journal of Knowledge Based Systems, Vol.18, No.7, pp.327 -- 333 (Elsevier, 2006)

Yukio Ohsawa, Scenario Understanding of Hepatitis Progress and Recovery by Annnotation-based Integration of Data based Scenario Maps, GESTS International Trans. Computer Science and Engineering Vol.22, N0.1., pp.65-76 (2005)

Yukio Ohsawa, Data Crystallization: Chance Discovery Extended for Dealing with Unobservable Events, New Mathematics and Natural Computation Vol.1, No.3, pp.373 - 392 (2005)

Renate Fruchter, Yukio Ohsawa, and Naohiro Matsumura, Knowledge reuse through chance discovery from an enterprise design-build enterprise data store, New Mathematics and Natural Computation Vol.1 No.3, pp.393-406 (2005)

Noriyuki Kushiro, and Yukio Ohsawa, a A scenario acquisition method with multi-dimensional hearing and hierarchical accommodation process, New Mathematics and Natural Computation Vol.2, No.1, pp.101-113 (2006)

Xavier Llor, a David E. Goldberg, Yukio Ohsawa, et al, Innovation and Creativity support via Chance Discovery, Genetic Algorithms, New Mathematics and Natural Computation, Vol.2, No.1, pp.85-100 (2006)

Yukio Ohsawa, Naohiro Matsumura, Naoaki Okazaki Understanding Scenarios of Individual Patients of Hepatitis in Double Helical Process Involving KeyGraph and DSV, The Fourth IEEE International Workshop on Soft Computing as Transdisciplinary Science and Technology (WSTST05), Muroran, pp.456- 469 (2005)

Tsuneki Sakakibara, Yukio Ohsawa Knowledge Discovery Method by Gradual Increase of Target Baskets from Sparse Dataset The Fourth IEEE International Workshop on Soft Computing as Transdisciplinary Science and Technology (WSTST05), Muroran, pp.480- 489 (2005)

Yuichi Washida, Hiroshi Tamura, Yukio Ohsawa Examining Small World Problem Using KeyGraph The Fourth IEEE International Workshop on Soft Computing as Transdisciplinary Science and Technology (WSTST05), Muroran, pp.490- 500 (2005)

(7) Interactions: Please list:

(a) Participation/presentations at meetings, conferences, seminars, etc.

Yukio Ohsawa: "Data Crystallization: A Project Beyond Chance Discovery for Discovering Unobservable Events," Invited Talk in IEEE International Conference on Granular Computing, Beijin (CDROM, 2005)

Yukio Ohsawa: Plenary Lecture "Chance Discovery: Data-based Decision for Design and Business" International Workshop on Chance Discovery, Aletheia University, Taipei (2005)

Yukio Ohsawa: "Data Crystallization: A Project Beyond Chance Discovery for Discovering Unobservable Events" IEEE International Conference on Granular Computing, Beijin (2005)

Yuko Ohsawa: Designing Systems for Chance Discovery, The Fourth IEEE International Workshop on Soft Computing as Transdisciplinary Science and Technology, Plenary Lecture (2005)

Yukio Ohsawa, Takaichi Itoh, Data Crystallizer: Tool for Discovering Unobservable Events, 1st Annual Workshop on Rough Sets and Chance Discovery (RSCD) in conjunction with 8th Joint Conference on Information Sciences (JCIS 2005), Salt Lake City (2005)

Kazuhisa INABA and Yukio OHSAWA, Study on a Method for Supporting Scenario Extraction from Time Series Information, 1st Annual Workshop on Rough Sets and Chance Discovery (RSCD) in conjunction with 8th Joint Conference on Information Sciences (JCIS 2005), Salt Lake City (2005)

Kenichi HORIE and Yukio OHSAWA, Extracting High Quality Scenario for Consensus On New Specifications of Equipment, 1st Annual Workshop on Rough Sets and Chance Discovery (RSCD) in conjunction with 8th Joint Conference on Information Sciences (JCIS 2005), Salt Lake City (2005)

Yukio Ohsawa, Human-based Annotation of Data-based Scenario Flow on Scenario Map for Understanding Hepatitis Scenarios, Proc. KES Conference (2005)

Noriyuki Kushiro and Yukio Ohsawa, A Scenario Elicitation Method in Cooperation with Requirements Engineering and Chance Discovery, Proc. KES Conference (2005)

Calkin A.S. Montero, Yukio Ohsawa, Kenji Araki Modelling the Discovery of Critical Utterances, Proc. KES Conference (2005)

Ken-ichi Horie, Yukio Ohsawa, Extracting High Quality Scenario for Consensus on Specifications of New Products, Proc. KES Conference (2005)

(b) Describe cases where knowledge resulting from your effort is used, or will be used, in a technology application. **Not all research projects will have such cases, but please list any that have occurred.**

- Visualizing the data of patent lists of a company, with our method of data crystallization, enabled to see new technologies not yet existing in the world.

(8) New:

(a) List discoveries, inventions, or patent disclosures. (If none, report None.).

- The basic method of data crystallization, enabling to realize hidden leaders and hidden demands in the market.
- The advanced method of data crystallization, which we call human-interactive annealing.

Patent disclosures: **None**

(b) Complete the attached "DD Form 882, Report of Inventions and Subcontractors."

(9) Honors/Awards: List honors and awards received during the contract period, or emanating from the AOARD-supported research project.

- Young scientist award, from the Japanese Ministry of Education, Culture, Sports, Science and Technology (May 2005)

(10) Archival Documentation: This section should include a description of your work at a level of technical detail that you think to be appropriate. Submission of reprints/preprints often satisfies this requirement. If you have questions on how to prepare this section, please discuss this matter with your AOARD program manager.

Attached (the copies of articles below)

Yoshiharu Maeno and Yukio Ohsawa, Understanding of dark events for harnessing risk, Chance Discovery for Real World Decision Making, Chapter 22m Springer Verlag (2006)

Yukio Ohsawa, Data Crystallization: Chance Discovery Extended for Dealing with Unobservable Events, New Mathematics and Natural Computation Vol.1, No.3, pp.373 - 392 (2005)

New Mathematics and Natural Computation

Data Crystallization: Chance Discovery Extended for Dealing with Unobservable Events *

Yukio Ohsawa †
School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, 113-8563, Japan
y.ohsawa@gmail.com

Received 17 June 2005

This paper introduces the concept of Chance Discovery, i.e., discovery of an event significant for decision making. Then, this paper also presents a current research project on Data Crystallization, which is an extension of Chance Discovery. The need for Data Crystallization is that only the observable part of the real world can be stored in data. For such scattered, i.e., incomplete and ill-structured data, data crystallizing aims at presenting the hidden structure among events including unobservable ones. This is realized with a tool which inserts dummy items, corresponding to unobservable but significant events, to the given data on past events. The existence of these unobservable events and their relations with other events are visualized with KeyGraph, showing events by nodes and their relations by links, on the data with inserted dummy items. This visualization is iterated with gradually increasing the number of links in the graph. This process is similar to the crystallization of snow with gradual decrease in the air temperature. For tuning the granularity level of structure to be visualized, this tool is integrated with human's process of chance discovery. This basic method is expected to be applicable for various real world domains where chance-discovery methods have been applied.

Keywords: Unobservable Events; Chance Discovery; Data Crystallization

1. Introduction

In this study, my research team is revealing events that are potentially important but have never been observed. Because they are not included in the data, existing mining methods hardly help in identifying such events. Data crystallization is the challenge to this difficult problem. It forms an extension of what we have been calling *Chance Discovery* since 2000 [1,2,3].

Chance discovery means the discovery of a *chance*, which is defined as an event significant for decision making. This has been a real challenge to go beyond the methodology of data mining, in that the new goal is the understanding of the

*This work was supported in part by the U.S. Government. Mr. Takaichi Ito, Keio University, contributed to this study as the software developer of data crystallization.
†School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8653 Japan (e-mail: y.ohsawa@ gmail.com).

meaning of rare events for making decisions, rather than learning rules for predicting these rare events [6,7]. For example, developers of cellular phone are seeking comments from users. Some comments significantly affect the decision of a developer to redesign cellular phones, so they can be regarded as "chances." Given these comments, data/text mining tools may be able to show the relations between comments, the similarities of users, etc. On the other hand, methods of chance discovery aid human-computer interactions to potentially achieve the detection of rare but influential events/words/items/people [8,9,10]. In order to realize Chance Discovery, we developed tools of data-visualization [11,12], to be coupled with human's perception of chances [13]. In the next section, we will review previous approaches to Chance Discovery.

2. The Problem of Chance Discovery

Let us define a *scenario* as a sequence of events and actions in a certain context. For example, suppose a customer of a drug store buys a number of items in series, a few items per month. He has an urge to do so because he has a certain persistent disease. In this case, fulfilling a remedy of the disease suggested by his doctor is the purpose covering the entire event-sequence, where an event is the patient's purchase of a drug. Here, the purpose to fulfill the remedy is the context covering the sequence. Then, this patient may learns about a new drug, and starts to take it for changing the scenario to a radical cure. After a month, his doctor gets upset hearing this change in the treatment due to the patient's ignorance regarding the risk of the new drug. Here, the doctor noticed the risky scenario in the context of side effects. The doctor urgently introduces surgical operation, a powerful method to overcome the side effects and change into the third scenario in the context of recovery.

In this example, we find two "chances" in the three scenarios. The first chance is the information about the new drug which changes from the first remedy scenario to the second scenario, i.e., the risky one. Then the doctor's surprise became the second chance to turn to the third scenario. According to the definition of "chance" by Ohsawa [1], i.e., an event or a situation significant for decision making, a chance occurs at the cross point of multiple scenarios as in the example above, because a decision is to select one scenario in the future. Based on this idea, methods of Chance Discovery may contribute significantly to sciences and business domains [3].

Here, let us stand on the position of a physician looking at the time series of symptoms during the progress of an individual patient's disease. The physician should take appropriate actions for curing this patient, at appropriate times.

$$\begin{aligned} Scenario1 &= event1 \rightarrow event2 \rightarrow event3 \ (the\ progress\ of\ the\ disease). \\ Scenario2 &= event4 \rightarrow event5 \rightarrow event6 \ (the\ effect\ of\ the\ new\ drug). \end{aligned} \quad (2.1)$$

Each event-sequence in Eq.(2.1) is a scenario as far as it is covered by some coherent context. For example, Scenario 1 is in the context of disease progression without treatment, and Scenario 2 is a scenario in the context of taking a new drug

with a side effect. Suppose there is another event 9, meaning the appearance of the new drug, shortly after event 2. The patient took this as a good chance, by just looking at the local relation among event 2, event 9, and event 4. For this patient's perception, the appearance of event 9 just after event 2 became essential for making a decision, and looked like a significant chance. However, the doctor looked at the overall relations among all events in the map in Fig. 1, and noticed the patient is going in a wrong direction. Thanks to his awareness of a side effect (event 5) of the new drug, he decides to perform a surgical operation.

Detecting an event at a cross point between multiple scenarios, such as event 2, event 9, and event 5 above, and selecting the scenario that includes such a cross point is the essence of Chance Discovery. In general, the meaning of a scenario with an explanatory context is easier to understand than an event shown alone. From Fig.1, we can understand the three basic scenarios, and the novel scenario emerging from connecting the basic scenarios via chance events. However, event 2, event 9, and event 5 as shown in Fig.1, are harder to understand if they are shown independently of other events. Without this understanding, it would be difficult to obtain the patient's consensus on introducing the surgical operation, because a rare event such as event 9 makes the situation harder to accept, and because this surgical operation itself is rare for ordinary patients.

For realizing such an understanding, visualizing the *scenario map* i.e. a two-dimensional graph on which user can find a meaningful scenario by finding a context covering a connected sequence of events, is useful. For example, on the scenario map in Fig.1, user can find the connected scenario beginning from Scenario 1, to move on via Scenario 2, and, finally, to reach Scenario 3. Here, we can regard each familiar scenario, such as Scenario 1 or Scenario 2, as an *island*. And, let us regard a path of links between islands as a *bridge*. In Chance Discovery, the problem then is to have the user obtaining bridges between islands, in order to explain the meaning of connections between islands by means of bridges, as a scenario which can be expressed in a language that is understandable for the user himself/herself.

3. The Human-Machine Interaction in Chance Discovery

In the prevalent term "scenario development," a scenario may sound like something to be "developed" by humans who consciously control the process by planning actions. However, valuable scenarios may often "emerge" unconsciously from communications of humans. For example, a *scenario workshop* developed by the Danish Board of Technology (2003) starts from scenarios of the future society that are preset by writers, then experts in the domain corresponding to the preset scenarios discuss scenarios for achieving further improvements. The discussants write down their opinions during the workshop, but it is rare that they notice all the reasons why those opinions came out and why the revised scenarios have been finally obtained.

This process of a scenario workshop can be compared with the KJ (Kawakita Jiro) method. In the KJ method, participants write down their initial ideas on

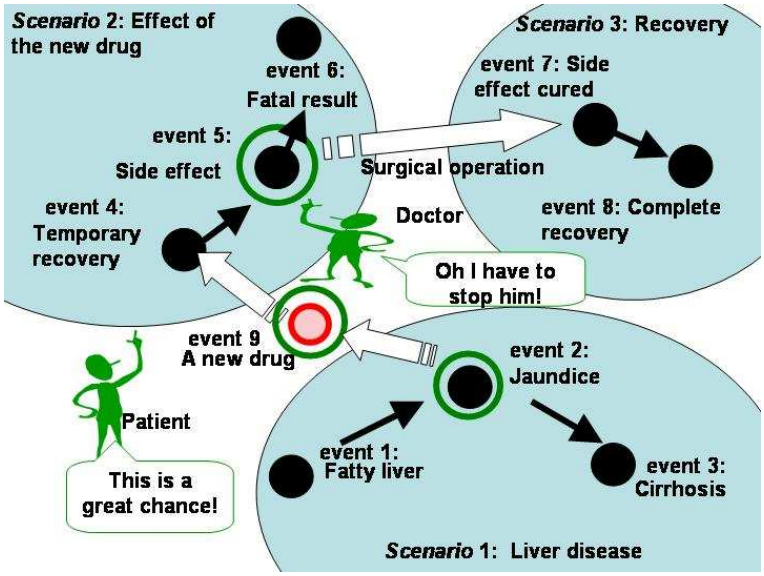

Fig. 1. A chance that exists at the cross point of scenarios. The scenario in the thick arrows emerged from Scenario 1 and Scenario 2.

KJ cards and hence arrange the cards in a 2D-space, in co-working for finding a good plan of actions. Here, the idea on each card reflects the future scenario in a participant's mind. The new combination of proposed scenarios, generated during the arrangement and the rearrangements of KJ cards, helps the emergence of new valuable scenarios. In some design processes, on the other hand, it has been pointed out that ambiguous information can trigger creations [4]. The common point among the scenario "workshop", the "combination" of ideas in the KJ method, and the "ambiguity" of the information to a designer is that scenarios presented from the viewpoint of each participant's environment, are bridged via ambiguous pieces of information about different mental worlds, which the participants attend. From these bridges, each participant indeed recognizes situations or events which may work as "chances" i.e., cross-over points for fusing others' scenarios with one's own. This can be extended to other domains than designing. In the example of Fig.1, the hopeful Scenario 3 after event 5 may be proposed by the doctor, and connected with Scenario 2 chosen by the patient before event 5. Here, event 5 played the role of cross-over point of the two scenarios, or the starting point of the thick arrow bridge.

In the studies of Chance Discovery, the discovery process has been supposed by Ohsawa to follow the Double Helix (DH) model [13] as shown in Fig.2 (Data Crystallization in Fig.2 is to be explained in later sections). The DH process starts from the initial state of the user's mind that is concerned with catching a new

chance. This concern is reflected to acquiring *external data* to be analyzed by a data-visualizing tool such as KeyGraph (to appear in later sections), which has been specifically designed for Chance Discovery. The visualization tool may depict each item in the data as a node, and the co-occurrence between items may be shown as links among nodes. Such a diagram has been regarded as a scenario map like Fig.1.

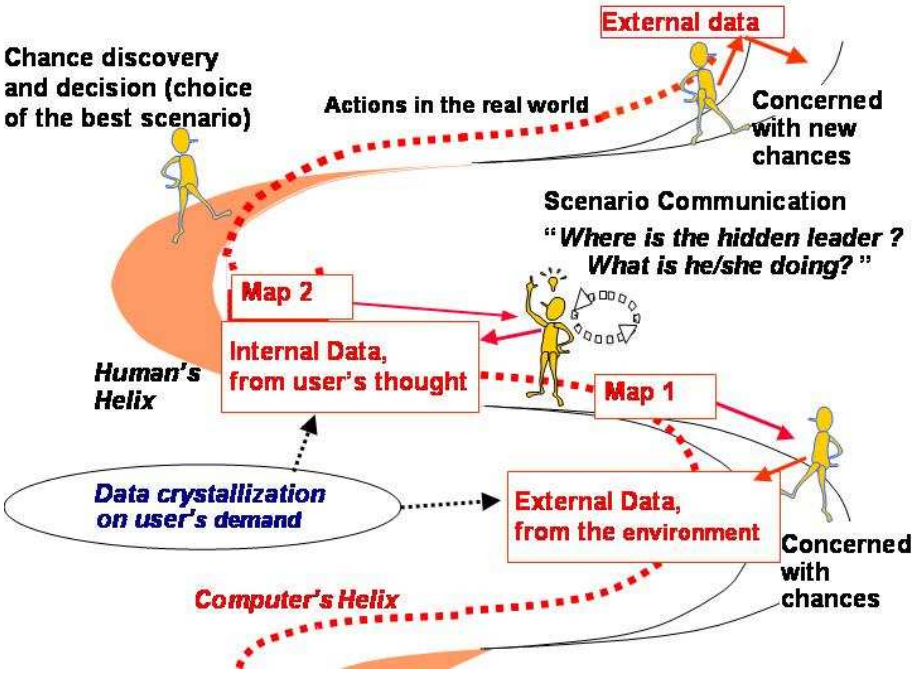

Fig. 2. Data crystallization on then double helix process.

Looking at the scenario map obtained, possible scenarios and their meanings emerge in each user's mind. Then, users participate in a co-working group for Chance Discovery, sharing the same scenario map. Here, they present the scenarios they find from the map. As a result, the computer acquires *internal data* i.e. the text data recording the thoughts and opinions presented in the discussion. The visualization tool is used now again: Words corresponding to contextual bridges are visualized, connected with prevalent daily-life contexts of participants. By this time, the participants discover chances on the bridges. Based on these chances, the users can make a new decision in the real world. Finally, the users perform a real action on which they obtain concerns with new chances, and the helical process returns to the initial step of the next cycle.

In the case of marketing, participants of a business meeting ran on the DH

process with sharing the result of KeyGraph. They looked at the map of their market using KeyGraph, where nodes correspond to products and links corresponding to co-occurrences between products in the customer's basket data. On this map, participants (market researchers) discussed with exchanging scenarios of customers living on various product-segments corresponding to local islands in the map. As a result, they found new scenarios of living customers who may buy products in all over the wide market. In contrast, previous methods of data-based marketing could identify focused segments of products and the scenarios in each local segment. This realized the hits of new products appearing in KeyGraph at bridges between islands. Thus, the participants of the DH process really discovered remarkable chances, and made real business profits [8].

4. Data Crystallization: A New Challenge

The complexity of the real world was sometimes beyond the reach of previous methods for Chance Discovery: A few nerd users of cellular phones, who do not send out comments frequently about their way of using cellular, are likely to create a new fashion causing strong influences on other users. The developer's question is "where is the innovative user?" If answers to these questions are available, the developer can continue to observe the behaviors of the innovative user, and may be able to catch the signs of new trends. This can be a significant chance in business, that may affect his decision.

It is meaningless to ask hundreds of monitors "who gave you the idea to use cellular phones in this way?" because users seldom see innovative users, but only see other users' accessories of cellular which are the indirect effects of the innovation. As a result, neither comments nor names of innovators can be included in the data on user's comments. Here arose the problem of Data Crystallization.

Data Crystallization, our new project that extends Chance Discovery, is dedicated to experts working in real domains where discoveries of unobservable events are desired. For example, let us consider intelligence analysis, where expert investigators of criminal-group behaviors are exploring links among members. The top leader (see the dark man at the top of Fig.3) of the criminal organization may phone a few times to sub-leaders managing local sections (Mr. A and Mr. B in Fig.3). For responding to these top-level commands, each local section holds its internal communication, via different media from that the top leader used for contacting sub-leaders. Then, the sub-leaders may meet to achieve consensus before responding to the top leader. Meanwhile, the leader does not appear in any meetings. In this way, someone never observed in meetings or mailing lists may be the actual leader.

5. The Method Overview of Data Crystallization

The objective of Data Crystallization is to detect (not only rare but) unobservable significant events. In this paper, I present an approach integrating two new methods,

to a breakthrough from the currents state of art in Chance Discovery.

The first is a method of visualizing data by inserting artificial dummy items. These dummy items mean unobservable events, of which the entities are totally unknown. The second is the human's process of discovery, where the chance may not be included in the data. For example, if the leader of a criminal group is unobservable, the intelligence analyst should become concerned with someone contacting sub-leaders moderating local meetings (Mr. A and Mr. B in Fig.3). Then, the analyst may move to the step of observing the living environments of Mr. A and Mr. B. In this way, human's interaction with the real world should be positioned in the process of data crystallization.

Basically, the presented method follows the Double Helix process as in Fig.2, which had been originally developed for Chance Discovery [13] and modified specifically for Data Crystallization. It begins with user's initial concern with occurring events which may be chances. On this concern, he/she collects data from the environment. The data are visualized in the computer-generated Map 1 of Fig.2, showing the computed relations between events in the real world, and the user begins to think of possible scenarios by connecting the events visualized. His/her thought here, or the communication of people working together, are stored in text. This text means stories rising from user's real-life experiences corresponding to the scenarios drawn in Map 1. This text is then visualized in Map 2. By looking at Map 2, possible scenarios composed of a sequence of events including unobservable chances become externalized. This lets the user become concerned with a certain part of the real environment, and brings the user to the start of the next cycle of the helical process. The effect of this process, to tuning the granularity of information about chances, enabled applications such as selling new products in marketing [8], detecting earthquake signs [14], treatment opportunity of hepatitis [9] etc. For Data Crystallization, we extend this process by putting the dummy-based visualization to Map 1 and Map 2. In this way, we aim at resolving harder problems than we challenged so far: Discovery of unobservable criminal leaders, revealing latent innovators, unobservable symptoms of hepatitis, unobservable active faults of earthquakes, etc.

6. KeyGraph: The Basic Tool for Visualizing Scenario Maps

KeyGraph [11,12] is a tool we had developed for visualizing relations among data items, corresponding to events in the real world. If the environment here means the society attacked by the teamwork of a criminal group, KeyGraph shows the relation of the group's members on the co-existing frequencies among members. In Eq.(6.2), let data $D1$ express a set of meetings, inserting a period (".") at each end of a meeting. Here, "member1" in Eq.(6.2) can be regarded as an event that a member appeared in a meeting place. Regarding each item in the data as an event rather than an object is meaningful in interpreting KeyGraph as a scenario map, where the sequence of events should be grasped from the connections between nodes.

$$
\begin{aligned}
D1 = \ & (set1)member1\ member2\ member3.\\
& (set2)member1\ member2\ member3\ member4.\\
& (set3)member4\ member5\ member7\ member6.\\
& (set4)member5\ member2\ member3\ member7\ member6.\\
& (set5)member1\ member2\ member7\ member6\ member9.\\
& (set6)member5\ member7\ member6\ member9. \qquad (6.2)
\end{aligned}
$$

KeyGraph takes the following steps, and is applied to data in the form of $D1$. Consequently, Fig.4 is obtained.

KeyGraph-Step 1: The M_1 most frequent items in the data (e.g., "member1" in Eq.(6.2)) are depicted with black nodes. The M_2 most strongly co-occurring item-pairs (i.e., the pairs of the highest values of the Jaccard co-efficient J in Eq.(6.3)) get linked via black lines.

$$J(X,Y) = p(X \cap Y)/p(X \cup Y). \qquad (6.3)$$

Here, $p(X \cap Y)$ means the probability that both item X and item Y appear in the same lines in data (as in $D1$ in Eq.(6.2)). $p(X \cap Y)$ can be computed by dividing the number of lines including both X and Y by the number of all lines in the data. Similarly $p(X \cup Y)$ is defined to mean the probability that either item X or item Y appears in the same lines in data. For example, member1, member2, and member3 in Eq.(6.2) are connected with black lines in Fig.4. Each connected graph here forms one *island* implying a basic context of the belonging members' life.

KeyGraph-Step 2: The M_3 items co-occurring with islands in the map most strongly, i.e., X of the largest $key(X)$ in Eq.(6.4), are obtained as hubs. For example, $member9$ in Eq.(6.2) is obtained here as a hub.

$$key(X) = 1 - \Pi_{Y:each\ island}\{1 - J(X,Y)\}. \qquad (6.4)$$

That is, the strength here between item X and island Y is computed as Jaccard co-efficient, after changing the name of each item in an island into the name of the island, in the given data. For example, if $member1$ is included in the first island, so it is renamed into $island1$. If $member5$ is in the second island, it is renamed into $island2$, in $D1$. Then, the co-occurrence strength between $member9$ and $island1$ is computed on Eq.(6.3), and is used in Eq.(6.4). In the obtained result, a path of links connecting islands via hubs is called a bridge. If a hub is rarer than black nodes, it is colored in a different color (e.g. red or white) than black. We regard such a hub as a candidate of chance, because it can be meaningful for a decision to jump from an island to another island.

Fig.4 supports the generation of a scenario of criminal behaviors, such as the one below, by recollecting information about the members from explicit or implicit (tacit) knowledge of intelligence analysts.

> "$Member1, member2$, and $member3$ are working together. And, $member5, member6$, and $member7$ form another group. When they meet $member9$, $member9$ may give commands to both groups from a higher level of the organization."

The appearance of a bridging member can be a central topic in the analysts' communication about crimes, and aids user's finding of chance events or items.

Fig.5 is the KeyGraph, for $D2$ in Eq.(6.5), the internal data from a communication of intelligence analysts about the criminal group. Each word is regarded here as an event, and a message from one participant as an event-set (i.e., as one line in Eq.(6.2)). The large islands in Fig.5, i.e., {member1, member2, member3} and {member5, member6, member7} mean the two groups are familiar to the analysts. The bridges of "message" and "forwards" linked to member9 show that member9 can just forward messages from one group to the other. On the other hand, we also find in Fig.5 that member9 may be a leader if member4 is "supposed" to be the secretary. Mr. Z decided to check the personal data of member4, as the "other" candidate for being the leader. However, from Fig.5, Mr. X and Mr. Y should note that Mr. Z was "sure" that member4 is the secretary. They should now check why Mr. Z made such contradictory comments. He may be telling a lie, or maybe member 4 is usually behaving ambiguously. Thus the focus of uncertainty is detected, and data can be collected in order to increase the granularity of information about the uncertain member. It is potentially possible now to decide to perform a new action for intelligence analysis.

$$D2 = the\ following\ text: \tag{6.5}$$

> "Mr.X: member1, member2, and member3 are working together.
> Mr.Y: And, member5 and member7 also form another group. I do not know member4...
> Mr.Z: I guess member9 is the leader of the all group of member1, member2, member3, member5, member6, and member7. I am sure member4 is their secretary.
> Mr. X: I think member5, member6, and member7 are a group. But member9 forwards the message from member1, member2, and member3, to member5, member6, and member7.
> Mr. Y: Suppose member4 is a secretary, who other than member9 can be the leader??
> Mr. Z: Let me check the personal data of member4 again. "

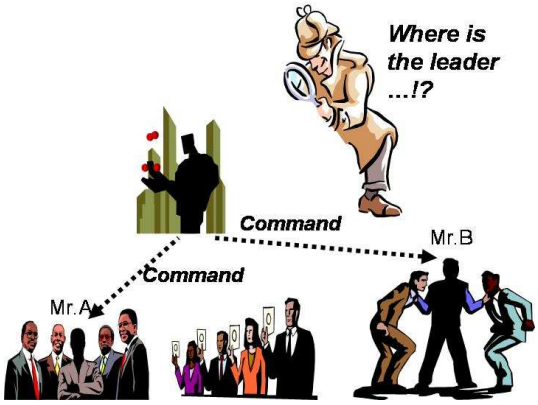

Fig. 3. Intelligence analysis seeking hidden leader.

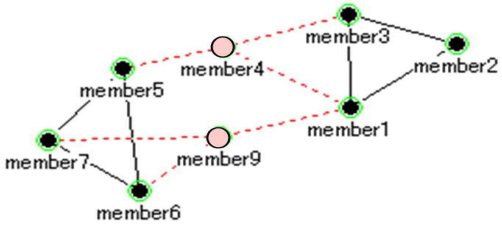

Fig. 4. An example of KeyGraph: Islands are obtained from $D1$ in Eq.(6.2), including sets {member1, member2, member3} and {member5, member6, member7} respectively. The nodes in and outside of the islands show frequent and rare items respectively, and member4 and member9 show rare hubs bridging islands.

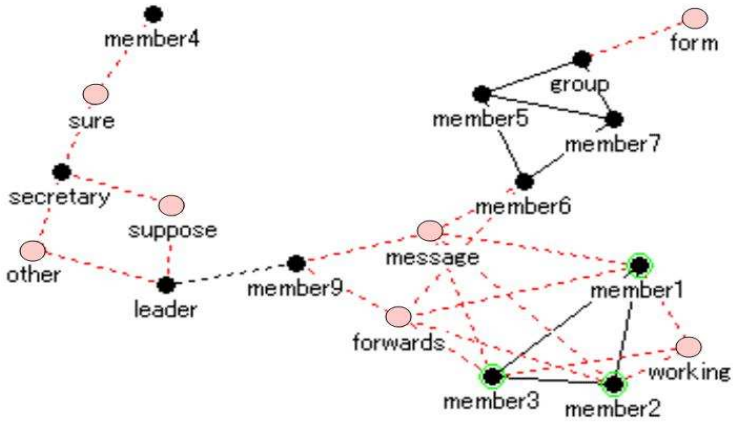

Fig. 5. KeyGraph, for the internal data. Islands are obtained from $D2$ in Eq.(4).

7. Data Crystallizer and The Data Crystallization Process

7.1. *Data Crystallizer: A Tool for Creating Dummy Items*

Data Crystallization aims at presenting the hidden structure among events including unobservable ones. This is realized on the process of Chance Discovery, with using a tool called Data Crystallizer, which inserts dummy items representing the potential existence of unobservable events, to the given data. Unobservable events and their relations with other events are to be visualized by applying KeyGraph, iteratively to the data, which were revised by inserting dummy items with Data Crystallizer. In each iteration, the size of each island is increased for reducing the granularity of the structure visualized. In essence, Data Crystallizer we developed runs the following procedure.

The procedure of data crystallizer

k :=1; $Hidden\_0$:= {}; $line\_0$:={}; M_1 := a value provided by the user;
for M_2 = 1 to M_1 $(M_1+1)/2$ do

for all $i, j \in 0, 1, \cdots, N$ such that $j \geq i$ do

if $line\_i$ and $line\_j$ are equal then $insert(D, k, i, j)$;

$H := keygraph(D, M_1, M_2, M_3 := M_1/2)$;
for j =1 to N do

If j $\notin H$ then $dlete\ (D, k, j)$;

If $H \neq Hidden\_k$ then

$k := k+1$;
$Hidden\_k := H$;
for m =0 to $k-1$ do

$delete(D, m, Hidden\_m \cap H)$;
$Hidden\_m := Hidden\_m \setminus H$;

Let me introduce the symbols employed: D is the data to be analyzed with KeyGraph in the function $KeyGraph(D, M_1, M_2, M_3)$. N is the number of lines (co-occurrence units) in the data, and $line\_j$ represents the set of items in the j-th line. H represents the set of line-numbers where the dummy items, which appeared on the bridges of the current KeyGraph, are positioned in the data. $Hidden\_i$ means the set of line-numbers with a dummy item which appeared on a bridge of the KeyGraph in the i-th level. The function $insert\ (D, k, i, j)$ means to insert k_j, the dummy node for the j-th line in the k-th level of crystallization, to the i-th line of data D and from data D. $delete(D, k, j)$ means to delete k_j, the dummy item for the j-th line on the k-th level, for all its appearances in data D.

Intuitively, we can explain the procedure as follows. Crystallization here means to present the structure of the relationship among items in and out of (dummy) the data. First, k, the level of crystallized structure, is set to 1. The value of M_1 (the number of black nodes in KeyGraph) is defined by the user(s). Then, M_2 (the

number of black lines) is incremented from 1, until all the nodes in the original data are connected and form a single island.

For each value of M_2, dummy items are inserted into D. The third and the forth lines of the procedure above mean: If 2 or more lines have the same set of items, the same dummy item is inserted to all those lines, suffixed with the line-number of the first of those lines. That is, k_j is inserted to the j-th line, and, if there is a line (the i-th line) of the same set of items as in the j-th line, k_j is inserted to all those lines.

To this data with inserted dummy nodes, KeyGraph is applied as in the fifth line. Then, the newest dummy items which did not appear on the bridges of KeyGraph are deleted from D as in the sixth and the seventh lines. The integer k, the level of crystallized structure, is incremented if H, the set of dummy nodes in the obtained KeyGraph, differs from $Hidden\_k$ i.e. the set of the latest dummy items obtained so far. If a line in the data includes 2 or more dummies, all the dummy items in the line except for the highest level are deleted, as in the eleventh to the thirteenth lines in the procedure.

After all, the following are obtained:

1) A new data set with dummy items, corresponding to hidden events that connect substructures in each level.
2) $keygraph(D,\ M_1,\ M_2,\ M_3)$ for the obtained data D, for arbitrarily determined values of M_1, M_2, and M_3. By increasing M_2, we can focus the output to the higher level of the hidden structure. By decreasing M_2, the granularity of the visualized structure is increased.

Data Crystallization works in the way like the crystallization of snow. A crystallizing item of the data plays a role like a particle of dust, which connects molecules of water in a cold temperature and forms a snow crystal. The increase in M_2 corresponds to the decrease in temperature, so the gradual increase in M_2 leads to a well-structured KeyGraph corresponding to a well-structured snow crystal obtained from gradual cooling of air.

7.2. *The Human-Machine Interaction in Data Crystallization*

The tool Data Crystallizer should work in Step 3) of the Double Helix peocess as described in the list below, because Data Crystallization is a kind of Chance Discovery. That is, Data Crystallization serves the understanding of deep-level chance events, but the dummy items corresponding to these events cannot be understood if the user is still in an early stage of Chance Discovery. There is a risk of disturbing user's understanding if a too complex structure is shown to someone who seeks simple information. Thus, Data Crystallizer works only if the user is concerned with unobservable level of the structure:

The Refined DH process for Data Crystallization
Step 1) Express the user's (or the users group) own concern with a chance.

Step 2) Obtain the external data, i.e., the data from the target environment, relevant to the current concern.

Step 3) Propose scenarios from the thoughts of user(s) by looking at the scenario map, which is the result of visual data mining with a tool such as KeyGraph, applied to the external data obtained in Step 2. If the participants want to investigate unobservable levels of the structure, use Data Crystallizer. Otherwise use KeyGraph without inserting dummy items.

Step 4) Visualize the internal data, i.e., the documented thoughts of user(s) in Step 3, by visual text mining.

Step 5) Choose the optimal scenario (by discovering chances if any), from the maps of Step 3 and Step 4.

Step 6) Evaluate the scenario obtained in Step 5) from the benefit/loss of the obtained scenario, and go to Step 1) if one obtains a new concern for improving the scenario.

8. A Running Case of Data Crystallization

We took a series of meetings in a faculty of 21 members, as the target data to analyze. In Da, a part of data on the participants are listed, obtained in Step 2) for our concern "where is the real leader ?" Here, each line corresponds to one meeting by some part of the faculty. Note that the names are arranged to hide real individual names, i.e., if reader finds a faculty of similar members, it might not be the case dealt with here.

$$\begin{aligned} Da \ = & \ tsubaki\ saru\ ogura\ kuwa \\ & tsubaki\ saru\ kuwa\ kawai \\ & kawai\ kuwa\ nagai \\ & ogura\ yoshida\ tsubaki\ kawai\ xu \\ & xu\ makimoto\ tsubaki\ yuji \\ & ryoke\ nagai \\ & \ldots \end{aligned} \tag{8.6}$$

Fig.6 is the result of KeyGraph in Step 3), for M_1=20, M_2=20, and M_3=20, from Da. Even though KeyGraph searched 20 hubs bridging between islands in this setting, we find all islands separated i.e., no bridges among them. That is, the faculty looked like a set of groups irrelevant to each other, in spite of the bridging function of KeyGraph. This was unreasonable, because the teamwork of this faculty was good enough to combine the knowledge of professors and make collaborative projects. Thus, we came to investigate deeper levels including hidden events. The dummy nodes are now inserted, denoted 1_x for the x-th line, to obtain Db below.

$$
\begin{aligned}
Db \; = & \; tsubaki \; saru \; ogura \; kuwa \; 1\_1 \\
& \; osawa \; yuji \; yoshida \; xu \; kawai \; sano \; 1\_2 \\
& \; tsubaki \; saru \; kuwa \; kawai \; 1\_3 \\
& \; kawai \; kuwa \; nagai \; 1\_4 \\
& \; ogura \; yoshida \; tsubaki \; kawai \; xu \; 1\_5 \\
& \; xu \; makimoto \; tsubaki \; yuji \; 1\_6 \\
& \; ryoke \; nagai 1\_7 \\
\ldots &
\end{aligned}
\tag{8.7}
$$

Fig.7 is the KeyGraph for *Db*. We now find that some dummy nodes remaining in the graph, forming the bridges among islands. For example, we find dummy 1_5 between yoshida and ogura. This means some hidden item relevant to the fifth meeting (the fifth line in Eq.(8.7)) made a significant bridge for the structure of the faculty. All dummy items which did not appear as bridges in Fig.7 are deleted from the data (see the sixth and the seventh lines in the procedure of Data Crystallizer).

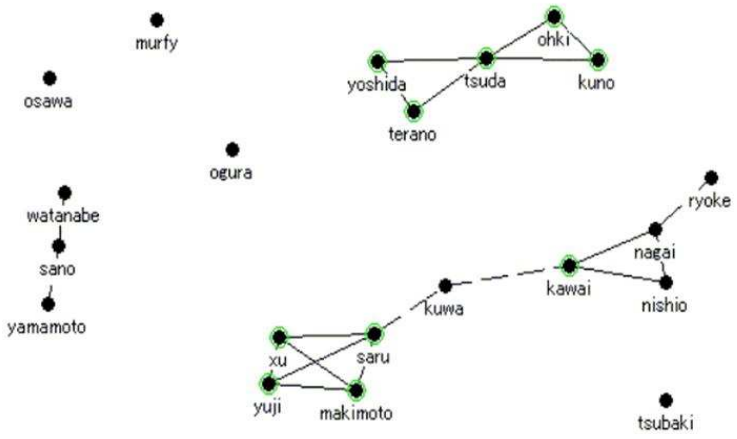

Fig. 6. The original KeyGraph for members of a group.

Then, new dummy nodes 2_x for the second level are inserted to obtain *Dc* in Eq.(8.8). However, let us skip the output of KeyGraph for *Dc* and just show the change in the data. That is, dummy nodes in the second level are deleted if they do not appear in the resultant KeyGraph, and the data change into *Dd* in Eq.(8.9). Having the tool run in this way to the third level, *De* as in Eq.(8.10) is obtained.

$$
\begin{aligned}
Dc = \ & tsubakisaruogurakuwa1\_12\_1 \\
& osawayujiyoshidaxukawaisano1\_22\_2 \\
& tsubakisarukuwakawai1\_32\_3 \\
& kawaikuwanagai2\_4 \\
& ogurayoshidatsubakikawaixu1\_52\_5 \\
& xumakimototsubakiyuji2\_6 \\
& ryokenagai1\_72\_7... \qquad (8.8)
\end{aligned}
$$

$$
\begin{aligned}
Dd = \ & tsubaki\ saru\ ogura\ kuwa\ 1\_1 \\
& osawa\ yuji\ yoshida\ xu\ kawai\ sano\ 2\_2 \\
& tsubaki\ saru\ kuwa\ kawai\ 1\_3 \\
& kawai\ kuwa\ nagai \\
& ogura\ yoshida\ tsubaki\ kawai\ xu\ 2\_5 \\
& xu\ makimoto\ tsubaki\ yuji \\
& ryoke\ nagai\ 1\_7 \\
& ryoke\ nagai\ tsubaki\ 1\_7... \qquad (8.9)
\end{aligned}
$$

$$
\begin{aligned}
De = \ & tsubaki\ saru\ ogura\ kuwa\ 1\_1 \\
& osawa\ yuji\ yoshida\ xu\ kawai\ sano\ 3\_2 \\
& tsubaki\ saru\ kuwa\ kawai\ 1\_3 \\
& kawai\ kuwa\ nagai \\
& ogura\ yoshida\ tsubaki\ kawai\ xu2\_5 \\
& xu\ makimoto\ tsubaki\ yuji \\
& ryoke\ nagai\ 1\_7 \\
& ryoke\ nagai\ tsubaki1\_7 \\
& ... \qquad (8.10)
\end{aligned}
$$

Fig.8 is the result for De, with M_2 increased up to 30. Increasing the number of black links (M_2) means to enlarge islands, for ignoring the local structure between small islands, and to focus attention on the higher level. Some dummy nodes in the same line appear in the same position in the graph, such as dummy 1_2 and dummy 3_2 in Fig.8. In such a case, only dummy 3_2 should remain here, so dummy 1_2 is deleted from the data set as in the tenth to the twelfth lines in the procedure of Data Crystallizer.

After obtaining De, the informative data with unobservable events, we can reduce the number of black lines, i.e., M_2, to obtain Fig.9 to see the lower-level

(dummy 1_x), the middle-level (dummy 2_x), and the high-level (dummy 3_x) structures of the human relations in the faculty. We apparently obtain newer findings than Fig.6. On Fig.9, the thoughts of some faculty members were collected as below.

- The 3_x dummy nodes represent the top level links. For example, Ogura was the head of the biggest department in the faculty two years ago, and his node is linked to the dean. Yoshida works in computer science, and is the current head of the department. Ogura and Yoshida are linked by 3_5.
- The next level (2_x) dummy nodes connect pairs e.g. {Ryoke, Nagai}, Watanabe, Sano. They were discussing the local arrangements of departments, i.e., middle-class management of the faculty.
- The next level (1_x) dummy nodes link pairs such as {Saru, Kuwa}. These correspond to proposals and acceptation from young staff such as Saru and Kuwa, i.e., bottom up proposals.
  (*continuing to other messages...*)

These messages constitute the internal data used in Step 4), in the Refined DH Process for Data Crystallization. By looking at Fig.8 obtained by KeyGraph for the internal data, the participants clearly became aware that the common interests of the dean (not included in the data of meeting participants), and the previous and the current heads of the biggest department are important for the management of the whole faculty. By looking at the common opinions of these heads, it is possible to detect signs of new trends of this faculty. In essence, the same prodecure as the one shown in this example is considered to be applicable to other human societies, such as criminal groups, consumers, researchers in a scientific domain, etc.

9. Conclusions

Data Crystallizing means to extend Chance Discovery to the discovery of significant events in more uncertain environment than we have been dealing with in studies on Chance Discovery. And, the sphere of real world applications linked from this basic research is expected to include intelligence analysis, development of new products, aiding corporate behaviors by detecting interest of employees, etc.

A relevant research area to Chance Discovery is Evidence Extraction and Link Discovery (EELD), where important links of people with other people and with their own actions are to be discovered from heterogeneous sources of data [13,14,15,16,17,18,19,20,21]. The difference between Chance Discovery and EELD, for the time being, is in the position of human factors in the research approaches. In Chance Discovery, the visualization techniques such as KeyGraph have been used for clarifying the effect of chances, by activating user's thoughts on scenarios in the real environment. On the other hand, the EELD program mainly contributed to identifying the most significant links among items more automatically and precisely than human.

Studies on EELD are coming to be oriented to coupling symbolic expressions of

human knowledge with a machine learning system [20], and also introducing the use of data visualization for decision making [17,18]. On the other hand, Chance Discovery has been integrating the human process of externalizing the tacit experiences with the power of machines for finding a surprising trigger to new actions in the real environment. That is, human's interaction with machine intelligence is coming to the centers of these two domains.

We finally predict the meeting point of Chance Discovery and EELD will be the detection of unobserved but significant events, as in the challenge of Data Crystallization. As shown in the jump from Fig.9 to Fig.10, the clarification of hidden links via unobservable events are finally up to the human thought. Human should look into more and more granular information about the environment, hand in hand with the crystallization of KeyGraph. This is like a scientist in a laboratory cooling the temperature slowly, carefully monitoring the experimental condition, in order to obtain a well-structured crystal.

References

1. Ohsawa, Y., McBurney, P. (eds), *Chance Discovery* (Springer Verlag, Heidelberg, 2003)
2. Abe, A., Ohsawa, Y. (eds), *Readings in Chance Discovery* (Advanced Knowledge International, Australia, 2005)
3. The Chance Discovery Consortium (CDC), Examples of Chance Discovery, http://www.chancediscovery.com (2004)
4. Gaver W.W., Beaver J., and Benford S., 2003, Ambiguity as a Resource for Design, in Proceedings of Computer Human Interactions
5. The Danish Board of Technology, 2003, European Participatory Technology Assessment: Participatory Methods in Technology Assessment and Technology Decision-Making,. http://www.tekno.dk/europta
6. Joshi, M., Kumar, V., Agarwal, R. Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements, *In Proc. of The First IEEE International Conference on Data Mining*, (San Jose, 2001)
7. Weiss, GM., and Hirsh, H (1998). Learning to Predict Rare Events in Event Sequences, *In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98),* (AAAI Press, Menlo Park, 1998) pp. 359 – 363
8. Ohsawa, Y., and Usui, M.,: Workshop with Touchable KeGraph Activating Textile Market, Abe, A and Ohsawa, Y (eds) Readings in Chance Discovery (Advanced Knowledge International, Australia, 2005) pp. 385 – 394
9. Ohsawa Y, Fujie H, Saiura A, Okazaki N, and Matsumura N, 2004, Process to Discovering Iron Decrease as Chance to Use Interferon to Hepatitis B, in Paton, R. (ed) Multidisciplinary Approaches to Theory in Medicine (Elsevier, The Netherland, 2005)
10. Ohsawa, Y., Soma, H., Matsuo, Y., Usui, M., and Matsumura, N., Featuring Web Communities based on Word Co-occurrence Structure of Communications, *Proceedings of the Eleventh Conf. World Wide Web (WWW11),* (ACM press, New York, 2002)
11. Ohsawa Y, 2003b, KeyGraph: Visualized Structure Among Event Clusters, in Ohsawa Y and McBurney P. eds, *Chance Discovery*, (Springer Verlag, 2003) pp. 262 – 275
12. Ohsawa, Y., Benson, N.E., and Yachida, M., KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor, *Proc. Advanced Digital*

Library Conference (IEEE ADL'98), (IEEE press, Los Alamos, 1998), pp. 12 – 18
13. Ohsawa, Y., 2003a, Modeling the Process of Chance Discovery, Ohsawa, Y. and McBurney eds, *Chance Discovery* (Springer Verlag, Heidelberg, 2003) pp. 2 – 15
14. Ohsawa, Y.: KeyGraph as Risk Explorer from Earthquake Sequence, *Journal of Contingencies and Crisis Management* 10 (3), pp. 119 – 128 (2002)
15. Senator, T., EELD Program, http://www.darpa.mil/ito/research/eeld/EELD_BAA.ppt (2001)
16. MeMeX: http://www.memex.com/le_i_analyst.html (2003)
17. Kovalerchuk, B., Relational Text Mining and Visualization, *Knowledge-Based Intelligent Information Engineering Systems and Allied Technologies (KES 2002)* (IOS Press, Amsterdam, 2002) , pp. 1549 – 1554
18. Kovalerchuk, B., Visualization and Decision-Making using Structural Information, *Proceedings of the International Conference on Imaging Science, Systems, and Technology (CISST'2001)* (2001) pp. 478 – 484
19. Oates, T., Schmill, M., and Cohen, P.R., Identifying Qualitatively Different Outcomes of Actions: Gaining Autonomy Through Learning . *Proc. of the International Conference on Autonomous Agents*, (ACM Press, New York, 2000) pp. 110 – 111
20. Sutton, C., Brendan, B., Morrison, C., and Cohen, P.R., Guided Incremental Construction of Belief Networks. *5th International Symposium on Intelligent Data Analysis* (2003)
21. Upal M.A., Performance Evaluation Metrics for Link Discovery Systems *Proceedings of the Third International Intelligent System Design & Applications*, (Springer-Verlag, New York, 2003) pp. 273 – 282

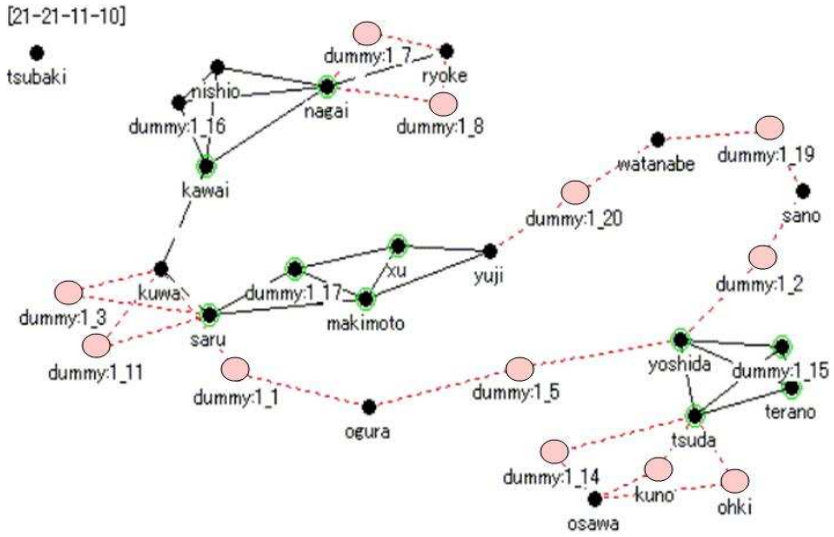

Fig. 7. The KeyGraph for data with first-order dummies (1_x).

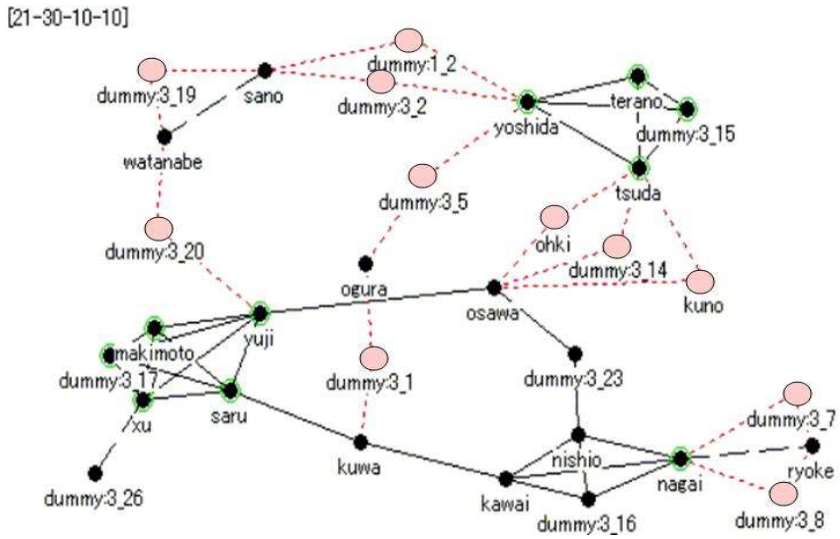

Fig. 8. The KeyGraph with third-order dummies, for M_2 =30.

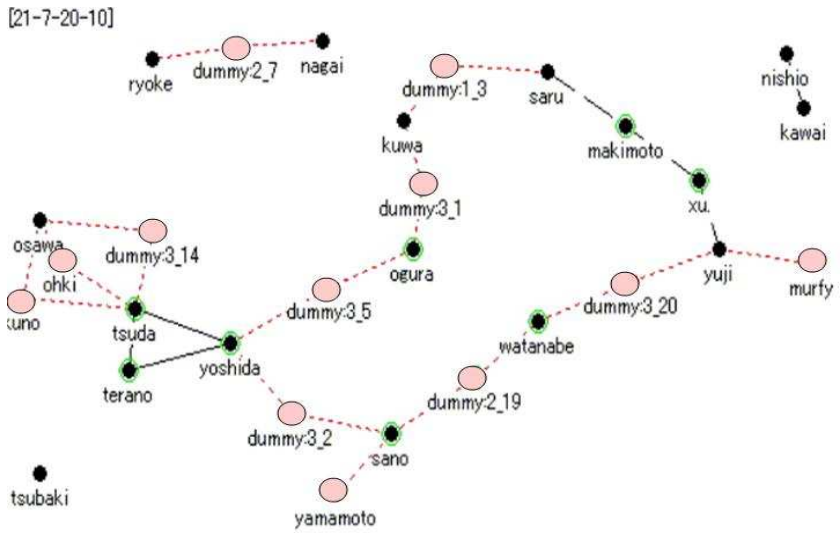

Fig. 9. The KeyGraph with third-order dummies, for M_2 reduced down to 7.

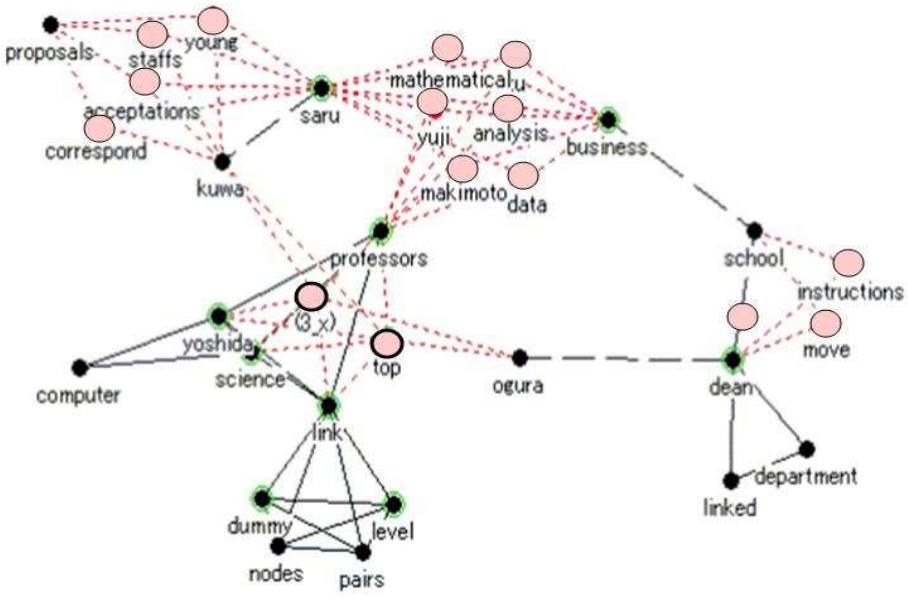

Fig. 10. The KeyGraph for comments on Fig.8.

Understanding of dark events for harnessing risk

Yoshiharu Maeno[1] and Yukio Ohsawa[2]

[1] Graduate School of Systems Management, University of Tsukuba,
3-29-11 Otsuka, Bunkyo-ku, Tokyo, 112-0012 Japan,
maeno@gssm.otsuka.tsukuba.ac.jp
[2] School of Engineering, University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8563 Japan,
ohsawa@q.t.u-tokyo.ac.jp

Abstract. There are invisible events which play an important role in the dynamics of visible events. Such an event is named a *dark event*. Understanding of the dark event is important for *harnessing risk in modern social and business problems*. A new technique has been deveoped to understand dark events and to extend the chance discovery process. The technique is *human-interactive annealing* for revealing latent structures along with the algorithm for discovering dark events. Test data generated from a scale-free network shows that the precision of the algorithm is up to 90%. An experiment on discovering an invisible leader hidden under an on-line decision-making circumstance and a trial for the analysis on unknown emerging technology are demonstrated.

1 Introduction

A chance means an event with significant impact on human's decision-making [Oh03a]. It could be conceived either as opportunity or as risk [Oh02]. The chance discovery process is designed for noticing a sign suggested by observed events and for putting new and significant scenarios into concrete shape [Oh03b]. In the process, a software tool named KeyGraph interfaces computational data processing, with human recognition and intuition. KeyGraph analyzes co-occurrence between observed events. It produces an event map and indicates a chance as a visual structure. The structure is a weak relationship bridging between multiple event clusters [Fu02]. In these features, the chance discovery is different from rare events or exception rules in data mining [Su05], [We98], and knowledge creation process [Ho94].

Experts of the chance discovery process, however, began to recognize a new problem, where the ordinary KeyGraph fail to visualize a latent structure hidden behind observation. It has been noticed empirically that important events composing the latent structure are neither visible nor observed in many social and business problems. Such invisible events are particularly important for harnessing risk. Let us describe two examples.

In human network analysis, it has drawn much attention to analyze terrorist organizations and to capture the signs of attacks. It is important to acquire information on leaders, close associates, important persons, and a chain of command

to the individual terrorists. The terrorist organizations hide such information in the visible data like communication logs, telephone records or emails. The leader seems to penetrate the organization like an invisible atmosphere and to synchronize individual terrorists toward the attack objective. This invisible atmosphere is a latent structure behind observed terrorist organization activities. Essentially, governments, intelligence offices, and secret services need understanding of the latent structure and an insight into a scenario for harnessing and removing risk from invisible terrorism.

In technology research and development, strategies on intellectual properties are critical to earning, costing, and even survival of companies. It is important to detect if competitor companies possess undisclosed surpassing technologies and expertise. Decision-making on making, buying, or licensing technologies is subject to such competitor companies' properties. Particularly, a sub-marine patent had been a great threat. Its publication is intentionally delayed by the applicant so that its presence and application can not be made visible. Such invisible technology is a latent structure behind complementarities and substitutability relationship among technologies and their holder companies. Essentially, strategists for corporate research and development need understanding of the latent structure and an insight into a scenario for harnessing and removing risk from hidden technologies.

From these examples, it is learned that there are invisible events which play an important role in the dynamics of visible events. Such invisible events are named dark events after dark matter in cosmology. New and significant scenarios for harnessing risk shall be put into concrete shape by understanding presence, nature, interaction and meaning of the dark events. But invisible dark events have not been within the scope of the chance discovery process. We have developed a new technique; *humna-interactive annealing* of latent structures along with crystallization algorithm of dark events to understand dark events. After studying the basic features of dark events, the principle of the technique and two application examples for harnessing risk in the real world are presented in the following sections.

2 Dark event

A new idea; *dark event* is introduced to formulate the problem described in section 1. The dark events are neither visible nor observable. Their associations to visible events form a latent structure hidden behind observation. But, the dark events are essential in the dynamics which governs temporal and spatial behavior, structure forming and life cycle of visible events. The dark event is analogous to dark matter in cosmology. The dark matter refers to hypothetical particles which do not emit or reflect radiation to be detected directly. But its presence can be inferred from gravitational effects on visible matter such as stars and galaxies. The dark matter hypothesis aims to explain several anomalous astronomical observations in the stellar dynamics. Estimates of the amount of the dark matter suggest that there is far more matter than is directly observable.

If dark matter does exist, it vastly outmasses the visible part of the universe. Before studying a means to analyze dark events closely, classification of events into four classes are presented. They are dark event, chance, visible event and event cluster. The chance, visible event and event cluster have been within the scope of the chance discovery process with KeyGraph.

- *Dark event*: The first class is dark event. The dark event is invisible because its occurrence frequency is very small. The dark event is diffusing randomly like an atmosphere because its association with other events is very weak. It does not tend to cling to a particular event cluster. It does not tend to appear as a pair with a particular event. In consequence, its co-occurrence is very small. This class of events has not been within the scope of chance discovery.

- *Chance*: The second class is chance. It is an infrequent but important event. Its occurrence frequency is very small. But its co-occurrence with a particular event or event cluster is not very small. KeyGraph are equipped with algorithms to analyze co-occurrence with Jaccard coefficient or Dependence coefficient. This class has been a major focus of chance discovery. KeyGraph visualize chance as a red node bridging between black node islands representing event clusters on an event map.

- *Visible event*: The third class is visible event. It is a frequent event. Its occurrence frequency is large. It can be observed easily. But its co-occurrence with a particular event or event cluster is not large. KeyGraph visualize a visible event as an isolated black node. So far the visible event has not been given large significance in chance discovery.

- *Event cluster*: The fourth class is event cluster. It is a set of frequent and strongly related events. Its occurrence frequency is large. Its co-occurrence with a particular event or event cluster is large as well. KeyGraph visualize an event cluster as a big black node island including many inter-connected black nodes. The event cluster is important as a reference point of observation to discover chance as a bridge node connected to it. The event clusters have a regular, ordered and stable nature.

The following is a working hypothesis on dark events and the evolution of chance. The dark events which are about to change into a chance may look like an emerging order in a chaotic structure. The chaotic structure close to the order may be discovered by identifying dense dark events and by analyzing them. On the contrary to the ordinary chance discovery process, human-interactive annealing of latent structures along with crystallization algorithm of dark events addresses the problem to understand dark events. Their details are described in the following sections.

- Hypothesis 1: Risk (or opportunity) shall originate in dense dark events, grows into a visible event (cluster), and matures into a well-understood scenario.

3 Annealing of latent structures

Before detailing the human-interactive annealing process, a little space is spent to learn a general meaning of *annealing*. Annealing in materials science is a heat treatment where the structure of a material is altered. It causes changes in the physical property such as strength through removal of crystal defects and the internal stresses. The annealing heats up a material piece until its temperature reaches a stress-relief point and cools down the piece slowly. Similarly, simulated annealing [Du00] is a probabilistic technique of computational optimization based on physical formulas describing the annealing in materials science. It is used to discover the optimal point in a large search space.

The human-interactive annealing similarly seeks the optimal point. It should be noted that the optimal point is in terms of human's creativity for new and significant scenarios. The annealing visualizes human recognition of the observed data into an event map. The optimal event map activates human's creativity most strongly. Our technique is based on the following working hypothesis on human recognition and creativity. The optimal event map is neither in ordered structure nor in chaotic random structure. The ordered structure is a group of well-understood concepts in human recognition. Mixing it with chaotic nature of dark events results in strong activation of human's creativity for new and significant scenarios. Such a structure is maintained in the basin of chaos between order and chaos [Ka96].

– Hypothesis 2: Mixing the ordered structure of well-understood concepts with chaotic nature of dark events shall result in strong activation of human's creativity.

The human-interactive annealing process is a combination of two complementary elements; crystallization algorithm on computers and human's interpretation. The two elements are illustrated in figure 1 with five event map examples. In the event maps, the event clusters and dark events are drawn schematically. The dark events are made visible, owing to the crystallization algorithm. The horizontal axis is the number of iteration. The vertical axis corresponds to the randomness of the visualized event structure. A parameter to control the randomness (like temperature) needs be introduced. It could be the number of event clusters or the total number of edges between events. The iteration is continued until human converges into complete understanding.

Crystallization algorithm is a breaking-through method by Ohsawa [Oh05], where dummy events which may potentially corresponds to the dark events are visualized. Yet, the complex algorithm and the complex graph obtained were hard to understand for users. It has been desired that user can reflect the user's interest in the visualization for focusing the obtained graph to understandable simplicity. We have modified the algorithm and incorporated it into the annealing process. In the crystallization, the computer analyzes the occurrence frequency and the co-occurrence of events. In the heating step, up to the specified peak temperature, the number of clusters and edges between visible events decrease.

Weak associations are destroyed. The crystallized dark events disappear. Then, a cooling step comes after the heating step, where event structures are solidified as temperature goes down. The number of crystallized dark events between clusters of visible events increases on an event map. The clusters are connected to each other to form a single large structure. The crystallization is followed by human's interpretation, where it is also checked whether the termination condition is fulfilled.

In the human's interpretation, let us assume that the process involves a group of humans, as in the previous cases of chance discovery. The humans put annotation to individual structures appeared on an event map, guess the meaning of dark events, and put scenarios into a concrete shape. If the structure does not match their intuitive recognition, they start annealing iteration again. It is a trigger of a heating step where event structures are dissolved as the temperature goes up to the next peak temperature. The peak temperature is specified based on the degree of understanding. When the understanding is poor, they should not change the peak temperature largely, but should stare at the current graph on an event map. On the other hand, if the structure matches their intuitive recognition approximately, they can re-start crystallization algorithm again to crystallize dark events further. If the structure implies novel scenarios of event occurrence finally, the iteration terminates, ending in complete understanding.

4 Crystallization algorithm

A new simplified crystallization algorithm has been developed to visualize dark events. This section details the algorithm, implementation with KeyGraph, and evaluation with measures of precision and recall. The basic idea of the crystallization algorithm is that visible dummy events are inserted to the input observation data to represent dark events. A dummy event is a symbolic expression of a latent structure containing dark events.

4.1 Crystallization of dark events

Observation data from which occurrence of events and co-occurrence between them can be evaluated shall be the input. For simplicity, we take basket data as an example of the input data format. The content of the basket is a set of events grouped under a specific subject. They may be a group of events observed simultaneously, or a group of events having some properties in common. Another typical input data format is vector representation of events in the multi-dimensional observation space. Before processing the baskets with the crystallization algorithm, the number of clusters, $|C|$ must be specified. At the first iteration, $|C|$ is initialized to be unity or a small number. After that, $|C|$ is gradually increased, based on the human interpretation. A generic crystallization algorithm under a specified number of clusters consists of five steps, event identification, clustering, dummy event insertion, co-occurrence calculation and topology analysis.

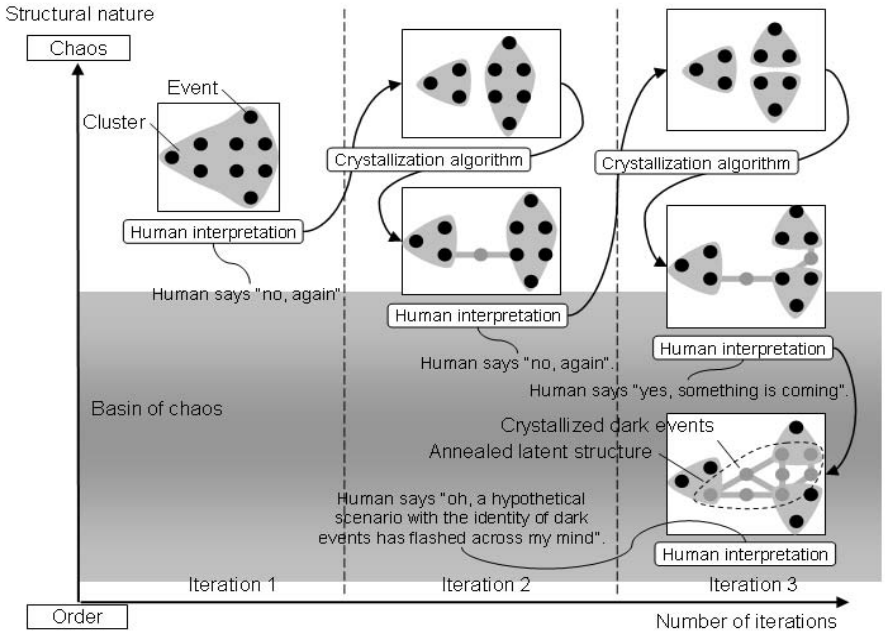

Fig. 1. Human-interactive annealing process for levealing a latent structure. The horizontal axis is the number of iteration. The vertical axis corresponds to the randomness of the visualized event structure.

1. *Event identification*: The all events appearing in the baskets $B = \{b_i\}$ ($i \in [0, |B| - 1]$) are picked up. The event set is denoted by E. The individual item is denoted by e_i ($i \in [0, |E| - 1]$).

2. *Clustering*: The event set E is classified into groups under a specified number of groups. This step can employ many existing technical expertise in statistics and machine learning such as clustering [Ha01], [Du00], unsupervised learning [Na05], projection of high dimensional data [Ag04], visualization [Hi99], and latent variable analysis [Bo89]. Clustering consists of partitioning a data set into subsets, so that the data in each subset share some similarity or proximity for some defined distance measure. Unsupervised learning is a method of machine learning where a model is fit to observations as input. It is distinguished from supervised learning by the fact that there is not a priori output to be learned or inferred from teacher data. The cluster set is denoted by C. The individual cluster is denoted by c_i ($i \in [0, |C| - 1]$).

   Existing clustering algorithms can be employed. Clustering may be hierarchical or non-hierarchical. The hierarchical clustering may be either divisive or agglomerative. The non-hierarchical clustering may use k-means algorithm, k-medoids algorithm, or equivalents. Kohonen's self-organization map (SOM) [Ko90], [Ha01], or graph theory based clustering methods [Du00] may also be applied. In either algorithm, a measure to evaluate similarity or dissimilarity between a pair of events is necessary. Similarity can be evaluated as co-occurrence of two items within baskets. Jaccard coefficient (equation (1)) and Dependence coefficient (equation (2)) are popular examples [Mu03], [Ma01]. The occurrence frequency of an event, e_i is denoted by $\mathrm{Freq}(e_i)$. They are an estimate of an association measure. The Dependence coefficient is called expected confidence, or lift.

$$\mathrm{Ja}(e_i, e_j) = \frac{\mathrm{Freq}(e_i \cap e_j)}{\mathrm{Freq}(e_i \cup e_j)} \tag{1}$$

$$\mathrm{Dep}(e_i, e_j) = \frac{\mathrm{Freq}(e_i \cap e_j)}{\mathrm{Freq}(e_i) \times \mathrm{Freq}(e_j)} \tag{2}$$

   Finally, calculated clusters c_i ($i \in [0, |C| - 1]$) are drawn on an event map. Links are drawn between a pair of events having large co-occurrence within individual clusters.

3. *Dummy event insertion*: A dummy event DE_i is inserted into a basket b_i [Oh05]. If $\{e_i\} \in b_i \equiv \{e_j\} \in b_j$ for $i \neq j$, DE_j is set to DE_i. The basket becomes $b_i \rightarrow \{\{e_i\}, \mathrm{DE}_i\}$. The dummy event represents a set of latent participants to the basket. It also corresponds to the subject to the basket. These are the first order dummy events. Higher order dummy events can also be inserted into baskets. For examples, the third order dummy event DE_{ijk} is inserted into baskets b_i, b_j and b_k.

4. *Co-occurrence calculation*: Co-occurrence between a dummy event and clusters is evaluated. In case of Jaccard coefficient, equation (3) is used. In equation (3), the function, max (maximal) may be replaced by functions, ave (average) or min (minimal), depending on the problem nature.

$$\mathrm{Co}(\mathrm{DE}_i, C) = \sum_{j=0}^{|C|-1} \max_{e_j \in c_j} \mathrm{Ja}(\mathrm{DE}_i, e_j) \tag{3}$$

   Two types of dummy events have large value of co-occurrence. One is those having large expected confidence with particular clusters. The other is those having relatively large expected confidence with relatively large number of clusters.

5. *Topology analysis*: The dummy events DE_i are ordered based on the co-occurrence with the clusters. The dummy events having large co-occurrence are picked up. The dummy events are connected to the clusters. The number of links between the dummy events and clusters is limited to 2 to 4 empirically. The number of picked up dummy events is increased until the all clusters are connected. Finally, the dummy events and links to the clusters are drawn on the event map. This structure reveals a latent structure consisting of dark events.

4.2 Implementation with KeyGraph

The crystallization algorithm can be implemented with the existing KeyGraph [Oh02]. KeyGraph employs a force-direct placement technique to draw a graph [Fu91]. The edges are replaced with a spring having characteristics depending on the co-occurrence to form a mechanical system [Su02]. An edge between vertexes having Jaccard coefficient above a threshold is subject to an attractive force. As a result, they tend to come close together. The vertices move until the mechanical system comes to an equilibrium state. Although the distance on the event map has no strict meaning, closeness between events approximately represents the strength of the relationship.

At first, dummy events are inserted to the original basket data. The first order dummy events are used. Higher order dummy events are neglected because their occurrence frequency ($\mathrm{Freq}(\mathrm{DE}_i) > 1$) results in wrong frequency analysis and clustering in KeyGraph algorithm. Then, KeyGraph output an event map. The number of black nodes is the same as the number of events $|E|$. The number of black links is a tuning parameter. The occurrence frequency of the dummy events is smaller than that of the original events. The dummy events do not appear as black nodes. The tuning parameter is adjusted to make the number of clusters C. The number of red nodes is zero. Finally, the number of red nodes is increased gradually so that the all black node clusters are connected. The dummy events become red nodes between the black node clusters.

4.3 Evaluation

We present a basic evaluation of the crystallization algorithm using test data generated from a scale-free network [Ba99]. The scale-free network is a commonly used model to describe human's communication, relationship or dependence in social problems. The scale-free network is suitable as a model for analyzing and harnessing risk. The scale-free network tends to contain centrally located hub events like leaders in an organization. The hub events influence the way the network operates. However, random deletion of events has little effect on the network's connectivity and effectiveness.

Figure 2 shows a scale-free network having 101 events. It includes a primary hub event (labeled 0-00) and five clusters (labeled 1-xx, 2-xx, 3-xx, 4-xx, and 5-xx). The clusters include secondary hub events (labeled 1-00, 2-00, 3-00, 4-00, and 5-00) and $95(= 19 \times 5)$ events. The event is connected with events in different clusters by the probability of 0.02. The occurrence frequency distribution of nodal degree is ruled by the power law; $y \propto x^{-2.7}$. The evaluation is for the crystallization algorithm rather than for the whole annealing process. Human's interpretation can not be applied because the scale-free network here does not have any understandable background context. The objective is to evaluate how much information regarding the primary hub event the crystallization algorithm can recover from the test data. The test data was generated in the two steps below.

- Step 1: One hundred basket data was generated from the scale-free network.
- Step 2: A latent structure regarding the primary hub event for the evaluation was configured to the basket data.

Events under a direct influence from an event are grouped into a basket. For example, we can imagine a situation where a person starts talking and a conversation takes place among neighboring persons. The area of such influence is specified approximately with the distance from an event. In this evaluation, we made up one hundred basket data consisting of events within two hops from an individual event in Figure 2. One hop is as long as one edge on the graph. Next, from the basket data, the primary hub event (0-00) was deleted so that the hub event was made invisible on the basket data. As a result, the primary hub event and the links inter-connecting the hub event and the five clusters became a latent structure hidden behind the basket data.

At first, we present a graphical result with a KeyGraph event map. Figure 3 shows an event map, resulting in 50 crystallized dummy events (pale bridges) inter-connected to 6 event clusters. The number of vertices in the clusters is still 100. The number of pale bridges was 50. Five large clusters correspond to the original 5 clusters in figure 2. Dummy events DE-35, DE-80, and DE-89 appeared between the two clusters. Thus, the basket data containing DE-35, DE-80, and DE-89 shall have additional relevant information on the latent structure. Actually, these basket data had contained the primary hub event before it was deleted. At least, three baskets were identified, from which we would obtain a clue regarding the invisible primary hub event. From these results, we confirmed

that basket data containing dummy events appearing as pale bridges between large event clusters indicate relevant information on the latent structure. The crystallization algorithm can recover information from the test data.

Next, we present quantitative performance evaluation to see whether the crystallization algorithm can output dummy events on the event map as a correct answer. In information retrieval, precision and recall have been used as evaluation criteria. Precision is the fraction of relevant data among the all data returned by search. Here, precision is evaluated by calculating the ratio of correct dummy events within all the dummy events emerging as pale bridges on the event map. The correct dummy events are those which were inserted to the basket data where the primary hub event had been deleted. In other words, they are those relevant to understanding the latent structure. Recall is the fraction of the all relevant data that is returned by the search among the all data. Recall is evaluated by calculating the ratio of correct dummy events emerging as pale bridges on the event map among the all correct dummy events. With precision and recall, we check whether the all dummy events and the only dummy events relevant to the primary hub event are picked up and visualized as pale bridges on the event map.

Figure 4 shows the calculated precision and recall as a function of the number of visible dummy events emerging as pale bridges. These results are under the same conditions as in figure 3 (six event clusters). The precision is 80% to 90%, when the number is less than 25. The first 25 dummy events correspond to the essential parts of the latent structure. It must be noted that the remaining 25 dummy events become noisier. This observation could be a heuristic rule to prioritize the dummy events to start analysis with.

5 Human interpretation

The human interpretation starts with putting annotation to clusters. Then, it proceeds to understand dummy events made visible by the crystallization algorithm. Some heuristic rules are referred to, to extract relevant areas from the event map.

5.1 Annotation

Annotation is additional information associated with a particular piece of data or a set of data in information. Annotation is a metadata including notes, comments, explanation, reminder or hints. It is useful to put annotations on the event map as a text in order to transfer one reader's interpretation to the other readers. Its principal function is, however, to convert the ambiguous awareness from intuition into an explicit and concrete understanding for the reader's own purposes.

The human interpretation starts with putting annotation to clusters. Clusters on an intuitively natural event map usually represent a single concept in human recognition. In other words, it constitutes a dimension in a human recognition

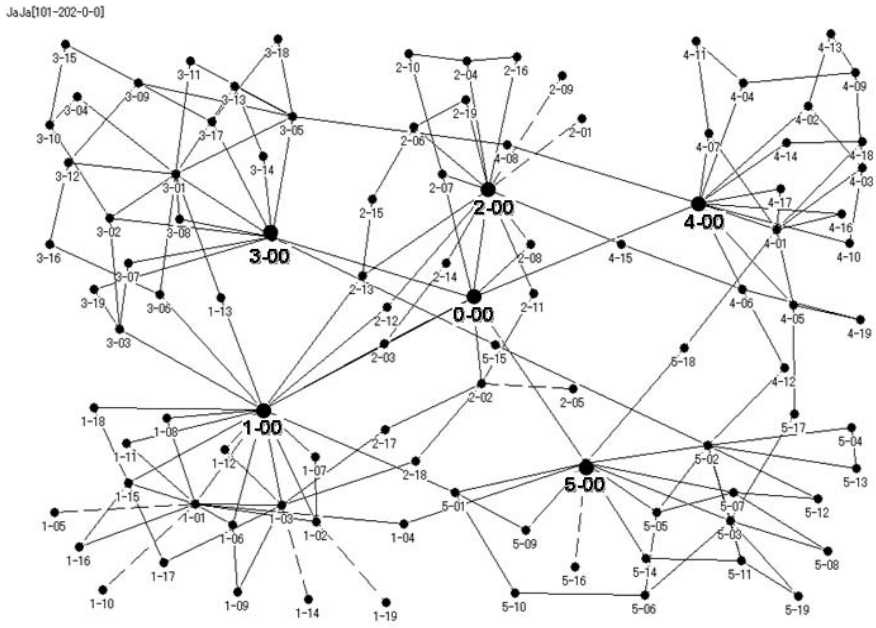

Fig. 2. Scale-free network with a primary hub event and five clusters. The clusters include secondary hub events and 95 events. The event is connected with events in different clusters at a probability of 0.02. The occurrence frequency distribution of nodal degree obeys the power law.

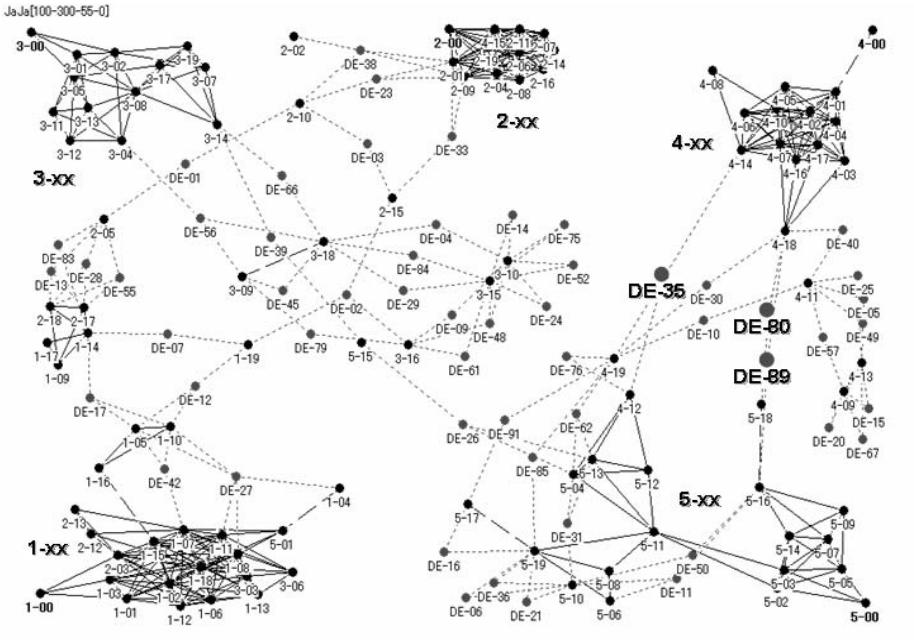

Fig. 3. The second iteration of the annealing process, resulting in fifty dummy events inter-connected to six event clusters.

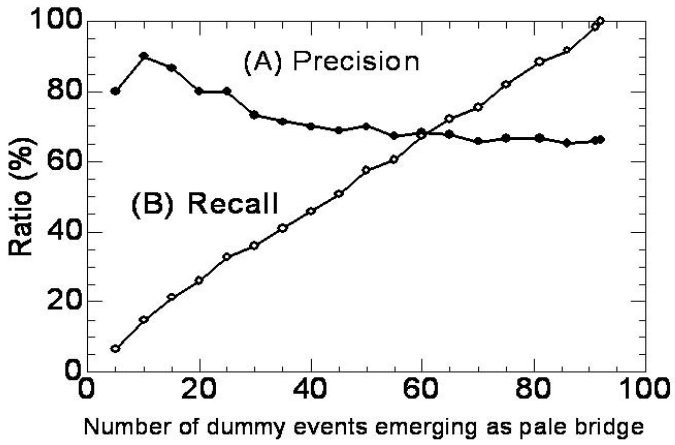

Fig. 4. Precision and recall of dummy events as a function of the number of visible dummy events under the same condition as in figure 3.

space. As the size of clusters increases, it gets easier to put annotation because larger clusters include more events and more information. Putting annotation from larger clusters to smaller clusters is a task to put aside human recognition and to configure the reader's own human recognition space. Next, human interpretation proceeds to understand dummy events on the event map. We need to know which dummy events to focus on initially. There are a few heuristic rules to start with. They are described next.

5.2 Heuristic rules for understanding

A heuristic rule is an empirical rule of thumb which usually produces a good solution or solves a simplified problem that contains the solution of complex problems. It often ignores whether the solution can be proven to be correct. But heuristic rule approach is effective when the problem is to complicate to define and treat mathematically, such as those in human knowledge, human recognition or human-computer interface. We have accumulated and confirmed some heuristic rules to extract a relevant structure from the event map after the annealing. The relevant structure the following heuristic rules indicate should be focused on to start investigation to imagine a scenario. It is also recommended to investigate the basket subject and content associated to the dummy events appearing in the focused structure.

– Heuristic rule 1: Imagine a scenario by carefully looking at the structure where many dummy events emerge as pale bridges between event clusters.

– Heuristic rule 2: Imagine a scenario by carefully looking at the structure where dummy events emerging as pale bridges are directly connected to a big event cluster.

Based on the heuristic rules, a generic density index to rank the importance of the latent structure has been derived. For individual gaps between clusters, the density index is the ratio of the number of dummy events to the distance between event clusters across the dummy events. The distance is the number of red nodes along the path from one cluster to another. Figure 5 illustrates the definition of the density index. According to the index, case (a) (index = 3/1) is more important than case (b) (index = 1/1). Case (b) is more important than case (c) (index = 1/2). The density index tells us to start investigating areas where the dark events are dense like in the case (a) and to understand the meaning of dark events in reference to the annotations put to the connected event clusters.

Relevant scenarios are lead by combining understood features of the dark events, problem specific knowledge, and experiences. Within the scenario, we shall get an insight into practical hypothesis beyond observation. The hypothesis may account for an influence from an unknown leader or a technology disruption by an unknown niche company. If the latent structure looks intuitively understandable, the human interpretation terminates the iteration in the annealing process.

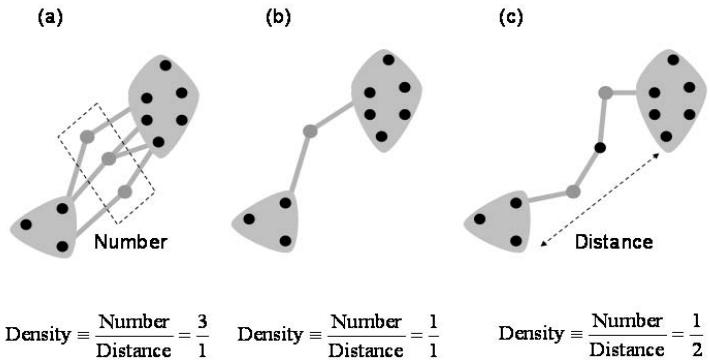

Fig. 5. Latent structures having different density index. The importance is evaluated with the ratio of the number of dummy events to the distance between event clusters. Case (a) is more important than case (b). Case (b) is more important than case (c).

6 Harnessing risk in the real world

Two demonstration is carried out to test the applicability of the annealing along with crystallization algorithm to the real world social and business problems. The first demonstration is an experiment on human network analysis. The latent structure is an invisible leader hidden in a mailing list for group based decision-making. The condition is similar to discovery a hidden leader in a terrorist organization. The second demonstration is an analysis on patents for technology research and development. The latent structure is an invisible emerging technological element. It is a trial for analysis on unknown emerging technology. Both are important examples in harnessing risk in the real world.

6.1 Discovery of an invisible leader

An experiment has been demonstrated to test the applicability of the whole human-interactive annealing process to social and business problems in the real world. The experiment is on human network analysis where we try to discover an invisible leader in a communication network with the annealing process. The latent structure is a chain of command from the invisible leader in a mailing list under a group-based collective decision-making circumstance. The invisible leader had a large influence on the discussion and opinions from individual members. A communication environment was prepared so that the invisible leader could instruct the individual members toward a favorable conclusion, orally without using the mailing list. During one month, 15 members participated in the mailing list, 220 emails were sent, and 56 basket data are observed. Subjects of the basket data are the titles of emails. The contents of each basket data are a set of members who sent and replied to the emails with the subject. They shall be the input to the annealing process. For example, a basket data contains a member initiating discussion by sending an email with "subject xyz" and members replying to the email with "re: subject xyz".

The result derived from the annealing of latent structures after the third iteration is shown in figure 6. Fourteen crystallized dummy events (pale bridges) become visible. They are inter-connected to seven-event clusters or isolated visible events. The figure includes the annotations put in human's interpretation. The annotation is based on the background knowledge on the problem and understanding of the member's characteristics. Four dummy events DE-07, DE-33, DE-35, and DE-45 appeared between a seven-member event cluster (Maeno, Ohsawa, Kushiro, Murata, Hashizume, Saito, and Murakami) and a single-member event (Horie). This area is important as the heuristic rules and density index evaluation of dark events in section 5.2 indicate. Table 1 shows the subjects and contents of the basket data including the four dummy members. Table 2 shows the actual commands from the invisible leader. Comparing the subject and email text with the detail of the actual commands, we confirmed that eight of twelve commands were successfully revealed by the four dummy events. From this analysis, precision is 100%(= 4/4) and recall is 67%(= 8/12). These eight commands seem more important than others in terms of an effort to converge

the discussion into conclusion. The annealing process accurately leads to the answer. Although the numbers 4 and 8 are small, this is a sound evidence of the performance, under the restriction where the invisible leader speak rarely. The experiment was successful in revealing the following two latent structures.

- *Instructions in Command*: The fact was as shown in table 2. The result was that the four subjects suggested by the dummy events were included in the commands. The annealing process revealed communication among the four dummy events representing the invisible leader and the members.

- *Chain of command*: The fact was that the invisible leader had sent commands primarily to members in a seven-member event cluster and a single-member event. The result was that the edges stemming from the four dummy events were along the commands. The annealing process revealed the chain of command from the invisible leader to the members. It is consistent with the annotations on observed characteristics of the members. From the intuitive observation, we were got convinced that the invisible leader should primarily contact with the three important members (Kushiro, Murata, and Horie) in the clusters.

Popular approaches in the present human network analysis are based on a network or graph theory. Scale-free networks [Ba99], or small worlds [Wa98] have been successful in describing many features of human activities and interactions. In addition to describing the human networks accurately, inferring a latent structure behind observation is getting more important. Such problem examples are the assessment of an organizational communication capability, the evaluation of human relational influence in a workplace, detection of collusion in a bid, and identification of disguise or aliasing in an Internet community. The human-interactive annealing along with the crystallization algorithm is expected to shed a new light on these problems.

6.2 Discovery of an invisible emerging technology

A simple trial for analysis on patents is demonstrated. Twenty nine patents applied in Japan are picked up as known technological expertise in the field of knowledge discovery. Patents provide with technological elements representing a measure to solve a specific engineering design problem. We try to identify an unknown but significant technological element by analyzing these patents. It may be a technology hidden by a rival company like a submarine patent, an emerging technology from other field of expertise, or a technology owned by a niche company or a small technician community. These latent strucutres are potential risk to corporate research and development. Subjects of the baskets are objective or preferred effect on the engineering design problems. Content of the baskets is a set of patent application numbers which is suitable for the subjects of the baskets. Thirteen baskets are configured. They shall be the input to the annealing process.

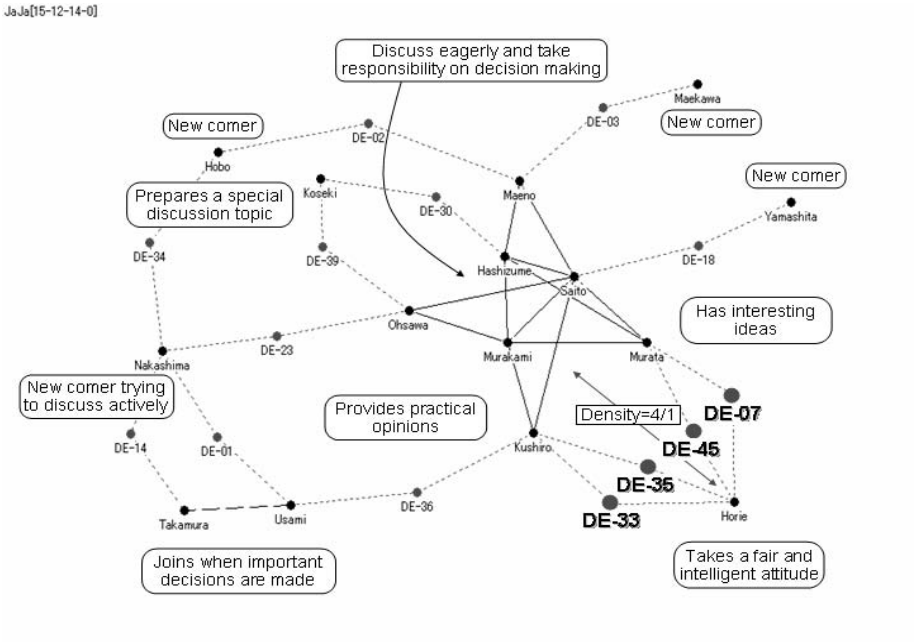

Fig. 6. Crystallized dummy events in the experiment with a mailing list to make a decision collectively under an invisible leader

Dummy event	Subject (email title)	Content (email sender and replier)
DE-07	Assign roles	Hashizume, Horie, Maeno, Murakami, Murata, Ohsawa
DE-33	Determine place	Hashizume, Horie, Kushiro, Maeno, Murakami, Saito
DE-35	Announcement on setup	Horie, Kushiro, Maeno, Murakami
DE-45	Voting on plans	Hashizume, Horie, Maeno, Murakami, Murata, Ohsawa, Saito

Table 1. Subjects and content of the basket including the four dummy members crystallized in figure 6.

No	Command from the invisible leader	Does it match the four dummy events ?
1	Announce about this mailing list	No
2	Invite new comers from outside	No
3	Introduce yourself	No
4	Make sub-groups to discuss individual topics	Yes (DE-07)
5	Play a role as a leader of a sub-group	Yes (DE-07)
6	Start discussion to assign tasks	Yes (DE-07)
7	Focus on particular subjects	No
8	Discuss on the place	Yes (DE-33)
9	Draw a conclusion on the recipe	Yes (DE-45)
10	Draw a conclusion on task assignment	Yes (DE-45)
11	Announce the arrangement	Yes (DE-35)
12	Announce the details	Yes (DE-35)

Table 2. Actual commands from the invidible leader to the members.

The result derived from the annealing of latent structures after the second iteration is shown in figure 7. Seven crystallized dummy events (pale bridges) become visible. They are inter-connected to eighteen-event clusters, smaller clusters, or isolated visible events. The figure includes the annotations put in human's interpretation. The annotation is based on the comments understood from the patents. Three dummy technological elements DE-01, DE-06, and DE-07 appeared between the biggest cluster and two two-event clusters. These areas are important as the heuristic rules and density index evaluation of dark events in section 5.2 indicate.

The biggest cluster corresponds to a set of conventional measures developed for statistical analysis or data mining in knowledge discovery. Particularly, discovery of association rules in knowledge discovery has evolved along three performance criteria. The first criterion is speed. This is required in real-time and on-line applications such as a contact center for product support and services. The second criterion is the amount of data. This is required in batch processing applications such as long-term customer trend analysis. The third criterion is quality. It means that more precise and more accurate association rules are required. The two two-event clusters incorporate technological elements for discovering unexpected knowledge and for visualizing knowledge respectively. Unexpected knowledge tends to be neglected in human's recognition, but significant for decision-making. In this sense, it is related to the chance discovery. Visualization is an important technical expertise which has been employed in many fields in science and engineering. These are mentioned as annotations in the figure.

The three dummy technological elements between the clusters suggest a new and unknown technological element which combines these three clusters. Here is the answer. The human-interactive annealing, about which you are reading, is just such a technology! It indicates unexpected risk by visualizing invisible dark events with use of the technical expertise in statistics and machine learning. The technological element represents a technique to incorporate human cognitive factor into the process. The result recommends the technology analyst to investigate closely whether potential competitor companies are developing such technological element or not. Although this analysis is for a simple demonstration purpose, it indicates how we should proceed to get an insight into a scenario for harnessing risk from hidden technological property based on a latent structure.

Popular approaches in the present technology research and development employ engineering design methods such as TRIZ (Theory of Inventive Problem Solving in Russian), Value Engineering (VE), or Taguchi method. These methods mainly aim at utilizing precedent successful cases and optimizing combination of technological elements under cost and quality constraint. Identifying an invisible new technological element emerging as a niche is getting more important. The annealing along with crystallization algorithm is expected to shed a new light on such problems.

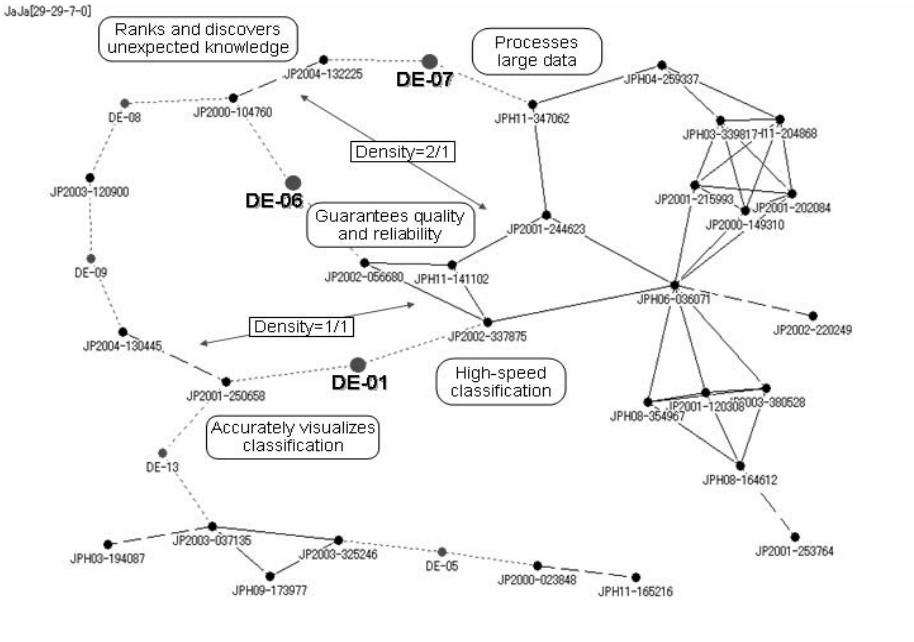

Fig. 7. Crystallized dummy events in the analysis on Japanese patents on knowledge discovery.

7 Summary

There are invisible events which play an important role in dynamics of visible events. Such events are named *dark events*. Understanding of the dark event is important for *harnessing risk in modern social and business problems*. Risk (or opportunity) may originate in dense dark events within a latent structure, grow into visible events or event clusters, and mature toward well-understood scenarios. To understand dark events, a new technique; *human-interactive annealing* of latent structures have been developed. The annealing process is combination and iteration of human interpretation and crystallization algorithm of dark events. Test data generated from a scale-free network showed that the precision of the algorithm is up to 90%. An experiment on discovering an invisible leader under an on-line collective decision-making circumstance was successful. The result indicates that we could discover a hidden terrorist leader and remove risk from the terrosist attacks. A trial for the analysis on patents for technology research and development were demonstrated. This could be a starting point for preparing for the impact from a hidden technology or an unknown emerging technology. The human-interactive annealing is a great advance in scenario writing where we shall get an insight into practical hypothesis beyond observation for harnessing risk in the real world.

Acknowledgement

The authors acknowledge Kiichi Itoh of Keio University for developing software. This work was partly supported by Asian Office of Aerospace Research and Development (AOARD), the US government.

References

[Ag04] C. C. Aggarwal: A human-computer interactive method for projected clustering, IEEE transactions on knowledge and data engineering, **16**, 448-460 (2004).

[Ba99] A. L. Barabasi, R. Albert, and H. Jeong: Mean-field theory for scale-free random networks, Physica A, **272**, 173-187 (1999).

[Bo89] K. A. Bollen: Structural equations with latent variables (Wiley series in probability and statistics). Wiley-Interscience (1989).

[Du00] R. O. Duda, P. E. Hart, and D. G. Stork: Pattern classification (2nd edition). Wiley-Interscience (2000).

[Fu02] H. Fukuda, and Y. Ohsawa: Chance discovery by stimulated groups of people: application to understanding consumption of rare food, Journal of contingencies and crisis management, **10**, 129-138 (2002).

[Fu91] T. M. J. Fruchterman and E. M. Reingold: Graph drawing by force-directed placement, Software - practice and experience, **18**, 1129-1164 (1991).

[Ha01] T. Hastie, R. Tibshirani, and J. Friedman: The elements of statistical learning: Data mining, inference, and prediction (Springer series in statistics). Springer-Verlag (2001).

[Hi99] Alexander Hinneburg, Daniel A. Keim and Markus Wawryniuk: HD-Eye: Visual Mining of High-Dimensional Data, IEEE Computer Graphics and Applications, September/October, 22-31 (1999).

[Ho94] Koichi Hori: A system for aiding creative concept formation, IEEE transactions on systems, man, and cybernetics, **24**, 882-894 (1994).

[Ka96] Stuart Kauffman: At Home in the Universe: The Search for Laws of Self-Organization and Complexity, Oxford University Press (1996).

[Ko90] T. Kohonen: The self-organizing map, Proceedings of the IEEE, **78**, 1464-1480 (1990).

[Ma01] N. Matsumura, Y. Ohsawa, and M. Ishizuka: Discovery of emerging topics by co-citation graph on the web, Proceeding of the fifth international conference on knowledge-based intelligent information engineering systems and allied technologies (KES), Osaka/Nara, Japan (2001).

[Mu03] T. Murata: Visualizing the structure of web communities based on data acquired from a search engine, IEEE transactions on industrial electronics, **50**, 860-866 (2003).

[Na05] I. T. Nabney, Y. Sun, P. Tino, and A. Kaban: Semisupervised learning of hierarchical latent trait model for data visualization, IEEE transactions on knowledge and data engineering, **17**, 384-400 (2005).

[Oh05] Y. Ohsawa: Data crystallization: chance discovery extended for dealing with unobservable events, New mathematics and natural computation, **1**, 373-392 (2005).

[Oh03a] Y. Ohsawa, and P. McBurney eds.: Chance discovery (Advanced information processing). Springer-Verlag (2003).

[Oh03b] Y. Ohsawa and Y. Nara: Decision process modeling across Internet and real world by double helical model of chance discovery, New generation computing, **21**, 109-121 (2003).

[Oh02] Y. Ohsawa: KeyGraph as risk explorer from earthquake sequence, Journal of contingencies and crisis management, **10**, 119-128 (2002).

[Su02] K. Sugiyama: Graph drawing and applications for software and knowledge engineers (Series on software engineering and knowledge engineering 11). World Scientific Publishing (2002).

[Su05] E. Suzuki, and J. M. Zytkow: Unified algorithm for undirected discovery of exception rules, International journal of intelligent systems, **20**, 673-691 (2005).

[Wa98] D. J. Watts, and S. H. Strogatz: Collective dynamics of small-world networks, Nature, **398**, 440-442 (1998).

[We98] G. M. Weiss, and H. Hirsh: Learning to predict rare events in event sequences, Proceedings of the fourth international conference on knowledge discovery and data mining (KDD), New York City, USA (1998).